REMARKS

Status of the Claims

Claims 125-132 and 148 are pending in the present application. Claims 133-147 are canceled. Claims 1-124 were previously canceled. Claims 126-132 are amended to correct antecedent basis. Claim 127 is further amended to correct an allegedly improper Markush group. The claims are amended without prejudice or disclaimer. Claim 148 is new. Support for new claim 148 is found throughout the application as originally filed, including, *e.g.*, in original claim 139. No new matter is entered by way of this amendment. Reconsideration is respectfully requested.

Informalities

The Examiner objects to claim 127 for allegedly describing an improper Markush group. The Examiner states that claim 127 should specify "selected from the group consisting of." The claim is amended according to the Examiner's suggestion. Accordingly, withdrawal of the objection is respectfully requested.

The Examiner further objects to claim 126-132 and 140-147 for specifying "[a] method according to." The Examiner suggests the claims be amended to specify "[t]he method according to."

Claims 140-147 are canceled. Accordingly, the rejection is moot in regard to these claims.

Claims 126-132 are amended according to the Examiner's recommendation. Accordingly, withdrawal of the objection is respectfully requested.

Issues Under 35 U.S.C. § 112, Second Paragraph

Claims 127 and 141 are rejected under 35 U.S.C. § 112, second paragraph, as allegedly indefinite. Specifically, the Examiner states that the instant claims specify a group of repetitive sequences and it is unclear whether the claims all require the repetitive sequences. The Examiner recommends amending the claims to recite a proper Markush group.

Claim 141 is canceled. Accordingly, the rejection is moot in regard to this claim.

Claim 127 is amended according to the Examiner's recommendation. Accordingly, withdrawal of the rejection is respectfully requested.

Issues Under 35 U.S.C. § 102

Fiegler

Claims 125-132, 139-141, and 145-147 are rejected under 35 U.S.C. § 102(a) as allegedly anticipated by Fiegler *et al.*, *Genes, Chromosomes & Cancer*, 36:361-374, ("Fiegler"). Specifically, the Examiner alleges that Fiegler describes the claimed methods. In support of this contention, the Examiner cites, *e.g.*, Fiegler at page 362, column 2, paragraph 3, page 363, column 1, paragraphs 1-2, and column 2, paragraphs 1-3, and page 364, column 1, paragraph 2. Applicants respectfully traverse.

Claims 139-141 and 145-147 are canceled. Accordingly, the rejection is moot in regard to these claims.

Notwithstanding the foregoing, Applicants submit that the Fiegler reference fails to describe all of the elements of instant independent claim 125. Specifically, Fiegler does not describe part (a) of claim 125, *i.e.*, randomly amplifying DNA from an isolated chromosome or part thereof, the amplified DNA being depleted of repetitive sequences and/or sequences that are over represented due to the random amplification. Contrary to the Examiner's assertion, Fiegler does not disclose these elements in the portions of Fiegler cited by the Examiner.

For example, the passage at page 362, column 2, paragraph 3, of Fiegler describes the arbitrary selection of BAC and PAC clones from which PCR products may be amplified. There is no disclosure in this section relating to repeat depletion of PCR products amplified from these BAC and PAC clones. The passage at page 363, column 1, paragraph 1, describes the isolation of genomic DNA from bacterial cultures that carry the BAC or PAC DNA. Again, there is no disclosure in this section relating to repeat depletion of PCR products amplified from BAC or PAC clones. Page 363, column 1, paragraph 2, and column 2, paragraphs 1 to 3, describe the DOP-PCR amplification of BAC and PAC clones and the application of the amplified DOP-PCR products to microarrays. However, there is no disclosure in any of these passages relating to repeat depletion of the amplified products which are applied to the array, (solid substrate). Page

364, column 1, paragraph 2, relates to the labeling of the test and reference DNAs for hybridization to the microarray and, accordingly, appears completely irrelevant to part (a) of instant claim 125.

Further, there is no suggestion that the DOP-PCR method used in Fiegler to prepare DNA for attachment to a solid substrate would lead to the depletion of repeat sequences in the amplified product. Applicants also submit that the inventor for this application, Nicole Hussey, states that, in general, DOP-PCR does not lead to a depletion of repeat sequences in an amplified product. In fact, as described below, Applicants submit that DOP-PCR amplification, at the very least, normally amplifies repeat regions and may actually be biased toward the amplification of repeat sequences.

For example, Pirker *et al.*, *Cytometry Part A*, 2004, 61:26-34, enclosed, describe that, with low amounts of template DNA, DOP-PCR excessively amplifies heterochromatic regions, *see* Figure 5C. In addition, Wang *et al.*, *Jpn. J. Hum. Genet.* 40(3):243-252, state that heterochromatin is largely made up of repeat sequences of the type depleted by Cot-1 DNA, *see* abstract only enclosed.

The scientific literature further reports analyses of non-mammalian DNA that show that DOP-PCR amplification, at the very least, amplifies repeat regions normally or biases toward their amplification. For example, Buzek *et al. Chromosome Res.*, 1997, 5:57-65, enclosed, demonstrated that clones from a plant species amplified by DOP-PCR were moderately to highly repetitive.

As a further example in relation to whole genome random amplification, Pinard *et al.*, *BMC Genomics*, 2006, 7:216-227, enclosed, showed extensive bias of DOP-PCR amplification toward repeat regions. Although this study was performed on bacterial chromosomes rather than human sequences, Pinard *et al.* clearly demonstrate that, in whole genome amplification, DOP-PCR does, in fact, amplify repeat sequences.

In view of the foregoing, Applicants submit that Fiegler does not disclose "amplified DNA being depleted of repetitive sequences and/or sequences that are over represented, due to the random amplification", as defined in part (a) of claim 125. In addition, Fiegler fails to make any suggestion that such depletion would be desirable. As such, independent claim 125 is not

anticipated by Fiegler. Dependent claims 126-132, which incorporate all of the elements of independent claims 125 are also not anticipated by the cited reference. Accordingly, Applicants respectfully request withdrawal of the rejection.

WO 00/24925

Claims 125-132, and 139-147 are rejected under 35 U.S.C. § 102(b) as allegedly anticipated by PCT Publication No. WO 00/24925 to Hussey *et al.*, ("Hussey"). The Examiner cites, *e.g.*, page 4, line 23-35, page 5, lines 1-14, and page 13, lines 18-27 of Hussey, as disclosing the features of part (a) of claim 125. The Examiner further cites page 6, lines 8-18, to support that WO 00/24925 discloses part (b) of instant claim 125.

Claims 139-147 are canceled. Accordingly, the rejection is moot in regard to these claims.

Applicants submit that Hussey fails to disclose the attachment of randomly amplified and repeat depleted nucleic acids to a solid substrate to function as target sequences in a CGH array, as described in claim 125. Applicants further submit that the sections of WO 00/24925 described by the Examiner relate to the random amplification of nucleic acids for hybridization to an immobilized nucleic acid, rather than to the immobilization of the amplified DNA on a solid substrate as a target for CGH.

As noted above, the Examiner cites page 4, lines 23-35, page 5, lines 1-14, and page 13 lines 18-27 of Hussey as disclosing the features of part (a) of claim 125. However, line 23 of page 4 clearly states that the DNA being amplified is the known chromosomal DNA and the unknown chromosomal DNA. In addition, the section starting at line 18 of page 13 of Hussey describes the amplification of DNA from a single cell and the subsequent hybridization of the amplified products to an immobilized metaphase spread.

The Examiner further cites page 6, lines 8-18, as allegedly disclosing immobilizing the amplified DNA to a solid substrate, as described in part (b) of claim 125. However, there is no disclosure in this passage of the immobilization of DNA, which has been randomly amplified from an isolated chromosome or part of an isolated chromosome, where the amplified DNA has been depleted of repetitive sequences and/or sequences that are over represented due to the random amplification, as defined in instant claim 125.

As stated above, parts (a) and (b) of claim 125 relate to the random amplification of DNA from an isolated chromosome or part of an isolated chromosome, wherein the amplified DNA has been depleted of repetitive sequences and/or sequences that are over represented due to the random amplification, and the subsequent immobilization of the amplified DNA on a solid substrate. The subsequent steps of claim 125 define the use of the immobilized DNA as a target for CGH, to which DNA amplified from a first and second sample may be hybridized. WO 00/24925 merely discloses amplification of DNA from cells, which is later hybridized to an immobilized nucleic acid. Furthermore, there is no disclosure in WO 00/24925, within the scope of part (a) of claim 125, which describes amplified DNA being immobilized to a solid substrate. Accordingly, Applicants submit that the disclosure of WO 00/024925 does not teach the elements of the instant claims. Further, this reference is not relevant to the method claimed in claim 125.

Based upon the foregoing, Applicants submit that the claims are not anticipated by WO 00/024925. Withdrawal of the rejection is respectfully requested.

CONCLUSION

In view of the above amendment and remarks, Applicants believes the pending application is in condition for allowance.

Should there be any outstanding matters that need to be resolved in the present application, the Examiner is respectfully requested to contact L. Parker Reg. No. 46,046 at the telephone number of the undersigned below, to conduct an interview in an effort to expedite prosecution in connection with the present application.

If necessary, the Commissioner is hereby authorized in this, concurrent, and future replies to charge payment or credit any overpayment to Deposit Account No. 02-2448 for any additional fees required under 37.C.F.R. §§1.16 or 1.17; particularly, extension of time fees.

Dated: July 20, 2009

Respectfully submitted,

By _____
Gerald M. Murphy, Jr.
Registration No.: 28,977
BIRCH, STEWART, KOLASCH & BIRCH, LLP
8110 Gatehouse Road
Suite 100 East
P.O. Box 747
Falls Church, Virginia 22040-0747
(703) 205-8000
Attorney for Applicant

Enclosures: Wang et al., Jpn. J. Hum. Genet. 40(3):243-252 (Abstract)
Buzek et al., Chromosome Res., 1997, 5:57-65
Pirker et al., Cytometry Part A, 2004, 61:26-34
Pinard et al., BMC Genomics, 2006, 7:216-227

A service of the U.S. National Library of Medicine and the National Institutes of Health

My NCBI [?] [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Go Clear Advanced Search

Limits Preview/Index History Clipboard Details

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: Jpn J Hum Genet. 1995 Sep;40(3):243-52.

Links

Cot-1 banding of human chromosomes using fluorescence in situ hybridization with Cy3 labeling.

Wang Y, Minoshima S, Shimizu N.

Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan.

We developed a new chromosome banding method by in situ hybridization of human Cot-1 DNA as a probe. Clear banding was produced on metaphase chromosomes of lymphoblastoid cells after probe detection with a fluorescent dye Cy3. Comparison with the known banding patterns revealed a similarity to the R-banding with some significant differences: some centromeric heterochromatin regions show Cot-1 positive bands. This suggests that some repetitive sequences from the heterochromatin regions constitute a major component of Cot-1 DNA. This unique chromosome banding method, Cot-1 banding, may be used as a supplement to the conventional karyotype analysis. Scanning analysis of the fluorescence intensities of Cot-1 banding and Q-banding are useful for objectively analyzing the banding pattern including a detection of chromosome aberrations. The Cot-1 banding with Cy3 is particularly powerful when applied for the gene mapping by fluorescence in situ hybridization (FISH) because red fluorescence of Cy3 for chromosome staining can be readily distinguished from green fluorescence of fluorescein isothiocyanate (FITC) for probe labeling. Using this novel method, we mapped a 4 kb-DNA fragment from myelin protein zero (MPZ) gene on the chromosome 1q22 to q23.

PMID: 8527798 [PubMed - indexed for MEDLINE]

Display AbstractPlus Show 20 Sort By Send to

Related articles

Toward a multicolor chromosome bar code for the entire human karyotype by fluorescence in situ hybridization [Genomics. 1997]

In situ hybridization banding of human chromosomes with Alu-PCR products: a simultaneous [Cytogenet Cell Genet. 1991]

Rapid fluorescence in situ hybridization with repetitive DNA probes: quantification by digital image analysis [Nat Genet. 1994]

Review [Extending the capabilities of human chromosome analysis: from high-resolution banding to [Cytogenet Cell Genet. 1997]

Review [Selective chromosome painting using in situ hybridization] [Sangre (Barc) 1993]

» See reviews... | » See all...

Recent Activity

Turn Off Clear

Cot-1 banding of human chromosomes using fluorescence in situ hybridization with Cy3

40[volume] AND 243[page]... (4) PubMed

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer

Chromosome Research 1997, 5, 57–65

Isolation and characterization of X chromosome-derived DNA sequences from a dioecious plant *Melandrium album*

Jiří Bůžek, Hana Koutníková, Andreas Houben, Karel Říha, Bohuslav Janoušek, Jiří Široký, Sarah Grant & Boris Vyskot

Received 18 September 1996; received in revised form 18 November 1996; accepted for publication by Pat Heslop Harrison 21 November 1996

A number of X chromosome DNA sequences have been isolated from a dioecious plant, *Melandrium album* (syn. *Silene latifolia*), using chromosome microdissection followed by degenerate oligonucleotide-primed polymerase chain reaction (DOP–PCR) amplification. Six DNA clones were selected and further characterized by DNA/DNA hybridization techniques to check their copy numbers, sex-specific methylation patterns, species specificity and positions on chromosomes. These clones were moderately to highly repetitive (approximately 10^3–10^5 copies per haploid genome) and none of them gave a positive signal on Northern blots. One of the clones yielded a sex-specific methylation pattern: its abundant non-methylated CCGG island was found only in males. All the clones also hybridized to two closely related dioecious *Melandrium* species (*M. rubrum* and *M. dicline*). Nucleotide sequences of two X-derived clones showed a number of internal short direct repeats; one of them strikingly resembled a plant conservative telomere sequence (TTTAGGG). None of the clones hybridized to the X chromosome only, but all were localized at the telomeric heterochromatic regions (DAPI C-bands) of both arms of a vast majority of *M. album* chromosomes using the fluorescence *in situ* hybridization (FISH) technique. However, the non-homologous arm of the Y chromosome (contrary to the arm homologous to the X chromosome, possessing the pseudoautosomal region) showed neither a DAPI C-banding-stained heterochromatin nor a FISH signal with any of the DNA probes tested, thus indicating its evolutionary diversification.

Key words: DNA methylation, *Melandrium album* (syn. *Silene latifolia*), sex chromosomes, subtelomeric heterochromatin

Introduction

Although many plant species are dioecious (approximately 5% of flowering plants), only a few of them possess well-defined heteromorphic sex chromosomes (for a review, see Westergaard 1958). The mechanisms of sex determination in various dioecious plants seem strikingly different, which indicates a relatively recent evolution of dioecy. Whereas in mammals, the Y chromosome is largely heterochromatic and the smallest in the genome, the Y chromosome in some dioecious plants is the largest one and, surprisingly, not heterochromatic (Charlesworth 1991, Grant *et al.* 1994). Generally, heteromorphic X and Y chromosomes possess two segments, non-homologous and homologous. The non-homologous segment may play a role in sex determination (as in the case of the Y chromosome in *Melandrium album*) and, as the distinct region between the X and Y chromosomes, it suppresses chiasma formation during the first meiotic prophase and metaphase. The homologous segment, or pseudoautosomal region, is subject to crossing over and chiasma formation.

Melandrium album (syn. *Silene latifolia*) belongs to the most popular and best-studied dioecious plant models. The genetic basis of its sex determination is well established. Males are heterogametic (XY); their Y chromosome harbours both male sex-determining and -promoting genes, and female sex-suppressing genes (Westergaard 1958). The pseudoautosomal region is restricted to a short segment at the end of one Y chromosome arm. The role of the X chromosome in sex determination and/or expression is ambiguous. Genetic analyses revealed that the X chromosome harbours a number of genes responsible for various phenotypical traits (Winge 1931). It was demonstrated that plants lacking this chromosome are unviable, so at least one copy of the X chromosome is necessary for the development of any sex (Ye *et al.* 1990). Our recent data indicate that one of the two female X chromosomes is hypermethylated and late replicating (Vyskot *et al.* 1993, Široký *et al.* 1994), thus resembling the dosage

J. Bůžek, H. Koutníková (present address: Institute of Genetics and Molecular and Cellular Biology, Strasbourg, France), K. Říha, B. Janoušek, J. Široký and B. Vyskot (corresponding author) are at the Institute of Biophysics, Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic. Fax: (+425) 41240500; Email: vyskot@ibp.cz. A. Houben (present address: Department of Genetics, University of Adelaide, Australia) is at the Institute of Plant Genetics and crop Plant Research, Gatersleben, Germany. S. Grant (present address: Department of Biology, University of North Carolina at Chapel Hill, USA) is at the Max-Planck-Institute for Plant Breeding, Cologne, Germany.

© *1997 Rapid Science Publishers*

compensation mechanism of X-linked genes as described in mammals (lyonization).

The fact that *M. album* sex chromosomes are easily distinguishable both from each other and from the autosomes offers a unique opportunity to construct sex chromosome libraries by flow sorting or microdissection (Grant *et al.* 1994, Veuskens *et al.* 1995). Other molecular biology strategies, such as random-amplified polymorphic DNA (RAPD) or representational difference analysis (RDA), were also used to select for Y-specific DNA sequences (Mulcahy *et al.* 1992, Donnison *et al.* 1996). Here, we describe a construction of the X chromosome DNA library using microdissection and DOP-PCR amplification to be used later to study X chromosome methylation and/or to screen for sex-determining genes in a female cDNA library. A number of DNA clones were isolated and six of them (giving a clear signal when hybridized with total genomic probe) were further characterized by Southern, Northern and fluorescence *in situ* hybridization.

Materials and methods

Plant material

Melandrium album Garcke (syn. *Silene latifolia*) plants were derived from an inbred line generated by eight generations of brother–sister mating (a gift from Dr J. van Brederode, State University, Utrecht, The Netherlands). *M. rubrum* (Weigel) Garcke, and *M. dicline* Willk. (syn. *Silene diclinis*) plants were kindly provided from a greenhouse collection by Dr J. Vagera (Institute of Experimental Botany, Olomouc, Czech Republic). To verify FISH hybridization signals on *M. album* of a different origin, plants from the collection of Dr J. Vagera and a male hairy-root cell line from Dr A. Mouras (University of Bordeaux II, France) were also used.

Chromosome preparation and microdissection

Root-tip meristems of *Melandrium album*, synchronized for cell divisions according to Pan *et al.* (1993), were fixed in 45% acetic acid. After washing in 75 mM KCl the meristematic tips were cut off and digested using an enzyme mixture consisting of 2.5% pectolyase Y-23 and 2.5% cellulase Onozuka R-10, diluted in 75 mM KCl and 7.5 mM EDTA, pH 4.0, at room temperature; about 30 root tips were harvested in 50 µl of enzyme solution. After 30 min maceration, the root tips were fixed in 45% acetic acid for 10 min and squashed on cover slips using the dry-ice technique.

A Zeiss Axiovert 35 inverted microscope equipped with a micromanipulator (Eppendorf) was used for chromosome isolation (microdissection). Twenty X chromosomes were collected in a 1-µl droplet containing proteinase K (0.5 mg/ml) in 10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.1% sodium dodecyl sulphate (SDS) and overlaid with water-saturated paraffin oil.

PCR amplification and DNA cloning

X chromosome microdissected DNA was amplified according to Pich *et al.* (1994) with a 6-MW degenerate primer [5′-CCG ACT CGA GNN NNN NAT GTG G-3′] (Telenius *et al.* 1992). PCR cycling involved a 4 min denaturing step at 95°C, five cycles each of denaturing at 94°C for 1 min, annealing at 30°C for 1.5 min, extension at 72°C for 3 min with a transition time of 3 min to 72°C. This was followed by 25 cycles at 94°C for 1 min, annealing at 55°C for 1 min, extension at 72°C for 1.5 min and a final extension at 72°C for 10 min. The PCR products were purified on Sephadex G50 columns, precipitated and checked by agarose electrophoresis. A DNA smear of 0.3–1.3 kb was detected. One-tenth of the PCR products was reamplified for 25 cycles under the conditions (annealing at 55°C) described above and cloned into pBluescriptKS$^+$ (Stratagene) using an internal *Xho*I recognition site derived from the primer and transformed into *E. coli* strain DH5α.

Extraction of plant DNA and RNA

Total plant DNA was isolated from leaves according to Dellaporta *et al.* (1983) followed by phenol/chloroform purification. Pooled female and male DNA samples were prepared from 10 plants of each sex (siblings). In order to estimate copy numbers of X-derived DNA clones, nuclear DNA from *M. album* female and male plants was prepared (Bedbrook 1981). Total RNA was extracted from the leaves of female *M. album* plants and from female and male flower buds using Trizol (Gibco-BRL) according to the manufacturer's protocol.

Southern and Northern hybridizations, DNA sequencing

To compare CG and CCG methylation within the CCGG recognition sites, total leaf DNA of *M. album* females and males were digested with an excess (12 U/µg DNA) of the restriction enzymes, *Hpa*II or *Msp*I, and the blots were hybridized with individual X-DNA clones. In order to find putatively X-linked species-specific DNA fragments, restriction patterns obtained after digestion with *Bam*HI, *Eco*RI and *Hind*III were compared in three related dioecious *Melandrium* species (*M. album*, *M. rubrum* and *M. dicline*). The restricted DNA samples were separated by agarose electrophoresis and blotted onto Hybond N (Amersham) membrane.

Individual plasmid inserts, the pool of the PCR-amplified fragments of X chromosomes or genomic *M. album* DNA, were labelled with [α-^{32}P]dCTP (Amersham) using the Decaprime kit (Ambion). Southern and Northern hybridization were performed as described by Sambrook *et al.* (1989) with a high-stringency washing step (0.2 × SSC, 0.1% SDS at 65°C, 2 × 30 min).

The completeness of restriction enzyme digestion of the total plant DNA was confirmed by rehybridization of membranes with a chloroplast DNA, pTB29, according to Fajkus & Fejch (1991). The pTB29 (Sugiura *et al.* 1986) was labelled with digoxigenin (DIG)-dUTP (Boehringer) and finally detected with anti-DIG–alkaline phosphatase conjugate and Lumigen PPD (Boehringer).

In order to estimate the copy numbers of the X–DNA clones in the *M. album* nuclear genome, the slot blot hybridization technique was used as described by Rivin (1986). Various amounts of denatured DNA (from 0.01 to 500 ng), both of the X-inserts and of the *M. album* nuclear DNA, were applied to membranes, hybridized with the respective radiolabelled X–DNA clones, and finally the spots were cut out and their radioactivity measured by scintillation counting. The copy

numbers of individual X-DNA clones were calculated based on the fact that the *M. album* diploid male nuclear genome consists of 5.645×10^8 bp, while the smaller female one has 5.532×10^8 bp (Vagera *et al.* 1994).

The X-clones were sequenced using [α-^{32}P]dATP (Amersham) and the DNA sequencing kit (Sequenase Version 2.0, USB) from T3 and T7 primers. Sequences were compared using the DNASIS (Hitachi) computer program.

Fluorescence *in situ* hybridization and chromosome banding

Seeds of *M. album* were sterilized, germinated in distilled water for 48 h, synchronized with aphidicoline (30 µM, 12 h) and metaphases were finally accumulated with oryzalin (15 µM, 4 h). Root tips were then enzymatically transformed into protoplasts according to Veuskens *et al.* (1995), fixed in Carnoy mixture and dropped onto slides. To follow meiotic pairing of X and Y chromosomes in *M. album* males, anthers from early flower buds were fixed in Carnoy mixture, squashed in 45% acetic acid and fixed onto slides using the dry-ice technique.

The X-DNA fragments were cut out from plasmids with *Xho*I and labelled with biotin-14-dATP using the BioNick kit (Gibco-BRL). In some experiments, an internal 2.478-kb (*Eco*RI) fragment of the tomato 25S-rRNA gene (Kiss *et al.* 1989) was also used to identify *M. album* autosomes. Slides were denatured in 70% formamide, 0.6 × SSC for 2 min at 70°C and dehydrated in ethanol series. The hybridization was performed according to Kenton *et al.* (1993). The biotin-labelled DNA duplexes were detected with fluorescein isothiocyanate (FITC)–avidin D and the signals were amplified by biotinylated goat anti-avidin and FITC–avidin D (Vector). The slides were mounted in Vectashield (Vector) containing 4',6-diamidino-2-phenylindole (DAPI) and propidium iodide (PI). The FISH signals were captured by a CCD camera and processed using the ISIS software (MetaSystems). The *in situ* hybridization signals with each X-DNA probe were checked on at least 20 mitotic plates.

Mitotic chromosome preparations were C-banded as follows: the slides were dehydrated in ethanol, air dried and immersed in a saturated solution of $Ba(OH)_2$ for 20 min at room temperature, washed thoroughly in water and incubated in 2 × SSC at 60°C for 5 min. The slides were then rinsed in phosphate-buffered saline (PBS), mounted in Vectashield containing DAPI and observed as described above.

Results

Construction of X chromosome library and selection of DNA clones

To construct an X chromosome library, the X chromosomes of *M. album* were microdissected, PCR amplified twice using the degenerate primer 6-MW and cloned into the pBluescriptKS$^+$ vector using the internal *Xho*I site within the primer. Forty-eight X-DNA clones were isolated and checked for their X chromosome origin by the hybridization with a probe composed of the pool of amplified X-DNA sequences. Twenty-seven clones showed a weak to strong hybridization signal of the fragment size between 0.2 kb and 1 kb. To estimate their abundance in the *M. album*, the cloned X-DNA inserts were hybridized with total female genomic DNA as a probe. For further analyses, two highly and three weakly hybridizing DNA clones were selected (X-12 and X-43 as the clones with a strong signal, and X-3, X-36 and X-41 giving a weak hybridization signal). Since the X-43 clone consisted of two *Xho*I fragments, it was further subcloned and analysed separately as X-43.1 and X-43.2.

Copy numbers and expression of selected X–DNA clones

The first analyses of the six X-derived representative DNA clones were aimed to estimate their copy numbers in female versus male *M. album* nuclear genomes. The hybridization signal was expected to be twice as strong in females as in males. However, the slot hybridization experiments did not show any significant differences between females and males. All six DNA clones tested were moderately to highly repetitive (from 7.1×10^3 to 1.4×10^5 copies per average haploid nuclear genome), thus representing <0.1% to 1.3% of the genome respectively (Table 1).

Northern blot analysis was performed to look for expression of the X-DNA clones. However, none of them gave a positive signal with RNA samples from female leaves or female and male flower buds.

Sex-specific DNA methylation patterns

Since the DNA clones tested were derived from the X chromosome that could be differentially methylated (Vyskot *et al.* 1993), we compared their CG and CCG methylation patterns in *M. album* males and females (Figure 1). The pooled genomic samples of females or males were cut with *Hpa*II or *Msp*I and hybridized with the individual X–DNA probes. All clones except X-3 yielded the same or a very similar type of hybridization pattern in females and males: strong high molecular weight signals when cut with *Hpa*II and dispersed patterns when cut with *Msp*I (Figure 1a). These data indicate a high degree of CG methylation in both females and males. The clone X-43.1 also displayed the same methylation patterns in both sexes and a heavy CG methylation (Figure 1b), but when the genomic DNAs were digested with *Msp*I, a ladder-like hybridization pattern typical of tandemly repeated DNA sequences appeared. In contrast to the X–DNA clones described above, prominent sex-specific methylation differences were observed when female and male genomic samples restricted with *Hpa*II or *Msp*I were hybridized with the X-3 probe (Figure 1c). The hybridization to high molecular weight fragments in female and male DNAs resembles the patterns obtained with the other probes showing a heavy CG and probably a moderate CCG methylation. However, the male DNA samples digested with either *Hpa*II or *Msp*I and hybridized with X-3 yielded a strong 530-bp fragment that was scarcely visible in females. To be sure that the genomic DNAs were cut with restriction

Table 1. Length, number of copies and estimation of CG methylation for each of the six X-derived DNA clones

DNA clone	Length (b)	Copy number per haploid genome	Fraction of nuclear genome (%)
X-3	254	7.1×10^3	< 0.1
X-12	750	4.3×10^4	1.2
X-36	310	1.5×10^4	0.2
X-41	256	9.1×10^3	0.1
X-43.1	208	2.3×10^5	1.3
X-43.2	400	2.3×10^4	0.4

enzymes completely, membranes were rehybridized with the chloroplast DNA clone (Figure 1d).

Species specificity of X-derived clones

A restriction pattern analysis comparing three different dioecious species of *Melandrium* was performed in order to find putatively X chromosome-linked species-specific bands that could subsequently be localized by genetic analysis. DNA samples from pools of female and male plants of *M. album*, *M. rubrum* and *M. dicline* cut with *Bam*HI, *Eco*RI and *Hin*dIII were compared after hybridization with the X-DNA clones. Both *M. rubrum* and *M. dicline* gave strong hybridization signals indicating that the *M. album* X-derived repetitive sequences were also present in these genomes. Entirely identical hybridization patterns were obtained after hybridization of DNA from the *Melandrium* species tested with the X-43.1 probe. This clone hybridized to a ladder-like pattern of DNA fragments (with individual units of 180 bp, Figure 2a) identical to the ladder of hybridizing *Msp*I-cut restriction fragments.

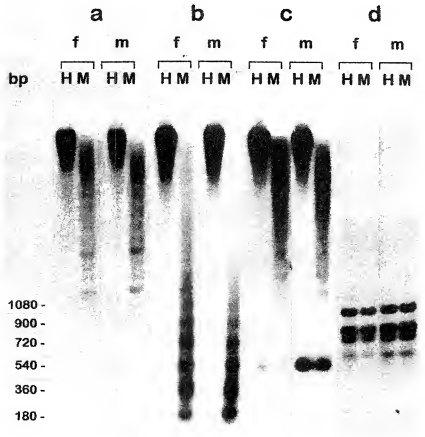

Figure 1. DNA methylation patterns (CG and CCG within the CCGG targets) of the pooled female (f) and male (m) *M. album* genomic samples cut with *Hpa*II (H) or *Msp*I (M) and hybridized with (a) X-12; (b) X-43.1; (c) X-3 and (d) the chloroplast probe pTB29 used as a control for complete digestion.

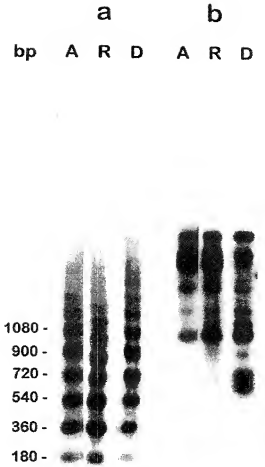

Figure 2. Restriction fragment polymorphism among related dioecious species, *M. album* (A), *M. rubrum* (R) and *M. dicline* (D), the pooled (female plus male) genomic DNA samples of which were cut with BamHI, EcoRI and HindIII, separated by gel electrophoresis and hybridized with (a) X-43.1 or (b) X-36.

This indicates that X-43.1 is a member of a highly conserved family of DNA sequences tandemly repeated in all three genomes. When X-36 (and X-3 to a lesser extent) were used as hybridization probes, some additional low molecular species-specific fragments were found in *M. dicline* (Figure 2b).

Nucleotide sequences of X-derived DNA clones

The DNA sequences of three clones (X-3, X-41 and X-43.1) showed a higher A+T content (52.3%, 52.4% and 60.0% respectively) but no significant mutual homologies (Figure 3). A search in the GenBank database revealed no homology to other DNA sequences using BLAST (the highest homologies correspond to AT-rich regions).

A striking feature of the X-3 sequence (254 bp) is a direct repeat of 18 bp (underlined with complete arrows in Figure 3). Within X-3, there is a *MspI/HpaII* restriction site and a GGCTCC sequence, which could represent a mutated *Bam*H1 site. The X-41 clone (256 bp) contains no internal repeats but several nucleotide sequences, from which *MspI/Hpa*II, *Bam*HI, *Eco*RI and *Hin*dIII restriction sites could arise by single base changes (these enzymes were used to study genomic DNA methylation and species-specific polymorphism). The clone X-43.1 (208 bp) appears to have the most interesting primary structure. Two classes of repeats are found within this fragment. The first class is formed by a tandem repeat represented by two complete units (22 bp) plus a fragment of a third one (7 bp) that are adjacent and in head-to-tail orientation (underlined with complete arrows in Figure 3). The second class consists of four imperfect 8-bp repeats that are scattered throughout this sequence (underlined with dashed arrows). Moreover, within X-43.1 two *MspI/Hpa*II restriction sites and two sequences resembling a degenerate monomer of the telomere sequence of *Arabidopsis thaliana* (TTTAGGG) were revealed (Figure 3).

In situ localization on mitotic and meiotic chromosomes

The hybridization sites of all six X-DNA probes tested were detected at heterochromatin blocks (DAPI C-bands) at the ends of *M. album* chromosomes (Figure 4). Hybridization signals of the individual probes differed in intensity, in accordance with their determined copy numbers, i.e. the signals generated by the X-43.1 probe were very strong (Figure 4a & b) and those of the probes, X-3 (Figure 4e) and X-41 (Figure 4f), were relatively weak. There were also clear differences in the size of heterochromatin blocks among individual chromosomes in mitotic metaphases revealed by both the C-banding and the intensity of hybridization signals. The most conspicuous heterochromatin blocks were found on the longer arm of the largest pair of autosomes and on the sex chromosomes (Figure 4a). On some chromosomes (usually on two or three pairs of smaller autosomes), the hybridization signals were sometimes missing or present on one chromatid only, which may reflect the small size of the corresponding terminal heterochromatin blocks.

Seven pairs of autosomes regularly showed a hybridization signal with each of the six DNA probes on both chromosome ends, four pairs on one end only. These four pairs were morphologically classified as the chromosomes bearing 25S-rDNA sites in accordance with Ciupercescu *et al.* (1990), i.e. chromosomes 5, 7, 9 and 10 (indicated in Figure 4a). The hybridization sites on chromosomes 5, 7 and 9 were located at the chromosome ends opposite to the 25S-rDNA sites. On

X−3

```
   1
  CCGTAGGAAG AGCCAGCGCT CTTCTTACGA TGTTTGTCGG AGGGGCGTGA
 51
  ATGCCCTGAT GTTGGACGCG TAAGTGCCTT CGATCCTTTT TTGATTACGT
101
  TTTCTTTTAT GTCTTTTATT TAGAAGTGAT CATAGGAATG ACCCCTTCCC
151
  GTTTAGAGCC TATCATAGGA ATGACCCCTC GTTTCTTATT TTAGTGGGTG
201
  CCTAGACCTT GGGAGAGCTA CCCTGAGGCT CCGGCCTTTG TGGCTGAGGT
251
  CCTT
```

X−41

```
   1
  CGTAATTTCT GTCGAGTAGG AATCCAGCAA GTCAGCAAAG TTACAAAGCG
 51
  GTTCAAGCCA CAATCACCGT GTCTCTCCAA GACGTTTATG CCATGGTCCT
101
  ACGTGTTCAA GCGTGGGCCG ATAATTGAAA GTCAGGATCT AAGATCGAGT
151
  CAGAGTGCGA GCCCGAGTCT TAGGCTTTCT ATCCCATAAT TGTTCTAGTA
201
  TCTTTCTTCG TGTCCATTTC CATTCCTAGT GACCACCTGG GGACTGTCCC
251
  ACGAGT
```

X−43.1

```
   1
  TATTAAGGGA GAATTTATTT CGGGACCCCG GAAACCTGAA AAACCGTGTA
 51
  TTATACATTT CAATTTCAAC CGTTCGGAAG GTCGTACCGG AACCTGTTTC
101
  TTTTTCTCCC TATTTTTCAA TCACGCGCCC CGAGTTCATT TCAGGAATCA
151
  ACCTGTTCGA TTTCAGGAAT CAACTTGTTC GATTTCAGCC TCAAATAACG
201
  AGCTTTAA
```

Figure 3. Nucleotide sequence organization of the three X-derived PCR clones. X-3 (% G+C=47.7), X-41 (% G+C=47.6) and X-43.1 (% G+C=40.0). The sequences are presented without the primer oligonucleotides. Bold letters indicate *Hpa*II/*Msp*I restriction sites (in X-3 and X 43.1); italics correspond to putatively point-mutated restriction sites for *Bam*HI, *Eco*RI or *Hin*dIII; underlined sequences with arrows represent direct repeats (in X-3 and X–43.1); and bold underlined sequences indicate the *Arabidopsis*-type degenerate telomeric sequences (in X–43.1).

chromosome 10, the hybridization signal was slightly subterminal and probably adjacent to the terminal 25S-rDNA site of this chromosome.

On the X and Y sex chromosomes, all six probes hybridized to the terminal heterochromatin blocks. These are located on both chromosome ends of the X chromosome but only on one end of the Y chromosome (Figure 4). As was shown by *in situ* hybridization on chromosomes at metaphase I of male meiosis (Figure 4c), the hybridizing end of the Y chromosome is located in the region homologous to the X chromosome. In Figure 4d, the pairing between the labelled chromosome ends of X and Y is clearly visible. Neither the DAPI C-banding nor the FISH analyses have shown the presence of terminal heterochromatin on the non-homologous arm of the Y chromosome. To test whether the presence of a heterochromatic block on only one arm of the Y chromosome was not caused by a chromosome rearrangement specific to this particular plant genotype (from the State University of Utrecht), C-banding and *in situ* hybridizations were also performed on mitotic chromosomes derived from *M. album* material provided by Dr J. Vagera and Dr A. Mouras with the same results (data not shown).

Discussion

The DOP-PCR protocol amplifies a broad range of target DNA sequences irrespective of their relative abundance. It has been demonstrated that this technique efficiently amplifies target DNA and, when total

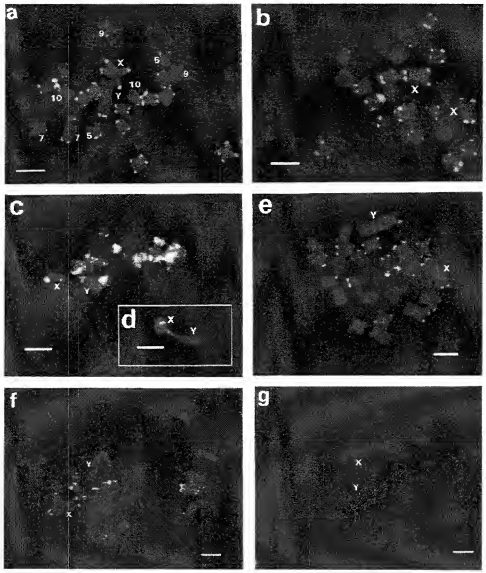

Figure 4. Fluorescence in situ hybridization and C-banding on *M. album* chromosomes. **a** Male metaphase chromosomes hybridized with X-43.1. Autosomes with the hybridization signal only at one chromosome end are marked by their numbers. **b** Female metaphase after hybridization with the same probe. **c** Meiotic chromosomes at late PMC metaphase I showing the X-43.1 signals. **d** Another example of X Y chromosome bivalent presented alongside. **e** Male metaphase hybridized with X-3. **f** Hybridization signals of X-41 on male metaphase. **g** C-banded male metaphase plate after DAPI staining. The X and Y sex chromosomes are indicated. Bars = 5 μm.

amplified DNA was used as a probe for fluorescence *in situ* hybridization, these DOP-PCR products hybridized uniformly over all chromosomes (Telenius *et al.* 1992, Pich *et al.* 1994). In our experiments, the DNA clones tested are evidently derived from only the telomeric regions of the X chromosome. This could result from sample bias, since a low number of clones was screened (48). Moreover, their successive selection (after Southern hybridization with the female genomic probe) could have favoured the identification of extremely abundant DNA sequences from the heterochromatic terminal region.

All selected X-derived DNA probes hybridized reproducibly to the same subtelomeric regions of a ma-

jority of the *M. album* chromosomes. The hybridization signals corresponded to the terminal heterochromatin as revealed by DAPI C-banding. In higher plants, the majority of C-bands corresponding to the heterochromatic regions are found at centromeres (e.g. *Pennisetum glaucum*, Kamm *et al.* 1994) or at telomeres (e.g. rye, Vershinin *et al.* 1995). In *M. album*, the heterochromatin regions are mainly located at the telomeric regions as shown by DAPI C-banding (Figure 4g). Telomeric regions (including the telomeric repeats and adjacent subtelomeric heterochromatin) are considered to have important functions in chromosome stability, chromatin behaviour during the cell cycle and meiotic pairing (Vershinin *et al.* 1995). In tobacco, a direct attachment of the telomeric $(TTTAGGG)_n$ repeat to a tandemly repeated sequence, HRS60, from terminal heterochromatin was observed (Fajkus *et al.* 1995). The juxtaposition of two different kinds of repeat sequences, telomeric and non-telomeric, in the X-43.1 clone suggests that there is a tight connection between telomeric and subtelomeric heterochromatin. We have observed that the six DNA clones tested were absent from the ends of chromosomes bearing the 25S-rDNA sites (except for chromosome 10). Here, the rDNA repetitive sequences forming the terminal nucleolus organizer regions (NORs) may be directly attached to the telomere repeats, as described on chromosomes 2 and 4 in *Arabidopsis thaliana* (Copenhaver & Pikaard 1996), and may substitute functionally for the subtelomeric repetitive sequences. A similar observation was made in some *Allieaceae* species lacking the *Arabidopsis*-type of telomeric repeats: their chromosome ends are formed by a highly repetitive satellite and/or rDNA sequences (Pich *et al.* 1996).

The Y-chromosome in *Melandrium* possesses the non-X-homologous (p) arm where the gynoecium suppressor and stamen initiation genes are located, and the X-homologous (q) arm that harbours anther maturation genes and also, on its end, the pseudoautosomal region corresponding to that on the q arm of the X chromosome (Westergaard 1958). In this work, we demonstrate that the pseudoautosomal region is, at least partially, composed of terminal heterochromatin. At metaphase I of male meiosis, the sex chromosomes are connected by their heterochromatic regions (Figure 4c & d). As presented by Irick (1994), heterochromatin is an essential part of the synaptonemal complex at meiosis in *Drosophila*. Albini & Schwarzacher (1992) showed that regions of rye chromosome, containing the heterochromatin-specific repetitive sequence pSc119.2, are included in synaptonemal complex at pachytene. These facts support the idea that heterochromatin plays an important role in meiotic chromosome pairing of higher eukaryotes in general. The presence of the heterochromatic block on only one (homologous) Y chromosome arm may thus be important for correct mutual orientation and pairing of *Melandrium* sex chromosomes at meiosis.

All three *Melandrium* species used in this study are dioecious and possess a pair of morphologically differentiated sex chromosomes. While in *M. album* and *M. rubrum* the Y chromosome is considerably longer than the X chromosome, in *M. dicline* the X and Y chromosomes are equal in length, but they differ in centromere position (Van Nigtevecht & Prentice 1985). Our selected X-derived *M. album* clones also hybridized to these related species; some differences in hybridization patterns of *M. dicline* may be connected with a divergent evolution of its sex (probably Y) chromosomes. There was no difference in the hybridization patterns of *M. album* and *M. rubrum* for any of the six probes tested. This is in agreement with a close evolutionary relationship between *M. album* and *M. rubrum* (Mastenbroek & Bredero de 1986).

The comparative DNA methylation studies on the female and male *M. album* genomes using the X-derived clones as probes were thought to reveal some differences resulting from the sex specificity and/or the putative dosage compensation of X-linked genes (Vyskot *et al.* 1993). However, none of the DNA clones tested was X chromosome-specific, so discussion of the dosage compensation hypothesis is irrelevant. Moreover, since no transcription of these six DNA clones was detected, the methylation differences found (both CG and CCG) cannot be attributed to their differential sex expression. It is probable, however, that the CCGG hypomethylated islands in males (detected with the X-3 probe) somehow reflect their positions near male-specific active gene(s) located on the Y chromosome and/or on autosome(s). The different hybridization patterns between the female and male genomes digested with *MspI* also confirm a frequent cytosine methylation at the outer position in 5′-CCG-3′trinucleotides as described in tobacco (Bezdek *et al.* 1992) and *Arabidopsis thaliana* (Jeddeloh & Richards 1996).

Acknowledgements

This research was supported by the Grant Agency of the Czech Republic (grants 521/96/1717 and 521/96/K117) and by the Grant Agency of the Czech Academy of Sciences (grant A5004601).

References

Albini SM, Schwarzacher T (1992) *In situ* localization of two repetitive DNA sequences to surface-spread pachytene chromosomes of rye. *Genome* **35**: 551–559.

Bedbrook J (1981) A plant nuclear DNA preparation procedure. *Plant Mol Biol Newsl* **2**: 24.

Bezdek M, Koukalová B, Kuhrová V, Vyskot B (1992) Differential sensitivity of CG and CCG DNA sequences to ethionine-induced hypomethylation of *Nicotiana tabacum*. *FEBS Lett* **300**: 268–270.

Charlesworth B (1991) The evolution of sex chromosomes. *Science* **251**: 1030–1033.

Ciupercescu DD, Veuskens J, Mouras A, Ye D, Briquet M,

Negrutiu I (1990) Karyotyping *Melandrium album*, a dioecious plant with heteromorphic sex chromosomes. *Genome* 33: 556–562.

Copenhaver GP, Pikaard CS (1996) RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosome 2 and 4. *Plant J* 9: 259–272.

Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: version II. *Plant Mol Biol Reporter* 1: 19–21.

Donnison IS, Siroky J, Vyskot B, Saedler H, Grant S (1996) Isolation of Y specific sequences from *Silene latifolia* and mapping of male sex determining genes using representational difference analysis. *Genetics* 144: 1891–1899.

Fajkus J, Reich J (1991) Evaluation of restriction endonuclease cleavage of plant nuclear DNA using contaminating chloroplast DNA. *Folia Biol* (Prague) 37: 224–226.

Fajkus J, Královics R, Kovarik A, Fajkusová L (1995) The telomeric sequence is directly attached to the HRS 60 subtelomeric tandem repeat in tobacco chromosomes. *FEBS Lett* 364: 33–35.

Grant S, Houben A, Vyskot B *et al.* (1994) Genetics of sex determination in flowering plants. *Dev Genet* 15: 214–230.

Irick H (1994) A new function for heterochromatin. *Chromosoma* 103: 1–3.

Jeddeloh JA, Richards EJ (1996) ^{m}CCG methylation in angiosperms. *Plant J* 9: 579–586.

Kamm A, Schmidt T, Heslop-Harrison JS (1994) Molecular and physical organization of highly repetitive, undermethylated DNA from *Pennisetum glaucum*. *Mol Gen Genet* 244: 420–425.

Kenton A, Parokonny AS, Gleba YY, Bennett MD (1993) Characterization of the *Nicotiana tabacum* L. genome by molecular cytogenetics. *Mol Gen Genet* 240: 159–169.

Kiss T, Kis M, Solomosy F (1989) Nucleotide sequence of a 25S rRNA gene from tomato. *Nucleic Acids Res* 17: 796.

Mastenbroek O, van Brederode J (1986) The possible evolution of *Silene pratensis* as deduced from present day variation patterns. *Biochem Syst Ecol* 14: 165–181.

Mulcahy DL, Weeden NF, Kesseli R, Caroll, SB (1992) DNA probes for the Y-chromosome of *Silene latifolia*, a dioecious angiospern. *Sex Plant Reprod* 5: 86–88.

Pan WH, Houben A, Schlegel R (1993) Highly effective cell synchronization in plant roots by hydroxyurea and amiprophos-methyl or colchicine. *Genome* 36: 387–390.

Pich U, Fuchs J, Schubert I (1996) How do *Alliaceae* stabilize their chromosome ends in the absence of TTTAGGG sequences? *Chrom Res* 4: 207–213.

Pich U, Houben A, Fuchs J, Meister A, Schubert I (1994) Utility of DNA amplified by degenerate oligonucleotide-primed PCR (DOP-PCR) from the total and defined chromosomal regions of the field bean. *Mol Gen Genet* 243: 173–177.

Rivin C (1986) Analyzing genome variation in plants. *Methods Enzymol* 118: 75–86.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual* New York: Cold Spring Harbor Laboratory Press.

Široký J, Janousek B, Mouras A, Vyskot B (1994) Replication pattern of sex chromosomes in *Melandrium album* female cells. *Hereditas* 120: 175–181.

Sugiura M, Shinozaki K, Zaita N, Kusuda M, Kumano M (1986) Clone bank of the tobacco (Nicotiana tabacum) chloroplast genome as a set of overlapping restriction endonuclease fragments: mapping of eleven ribosomal protein genes. *Plant Sci* 44: 211–216.

Telenius H, Carter NP, Bebb CE, Nordenskjüld M, Ponder BAJ, Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 13: 718–725.

Vagera J, Paulíková D, Doležel J (1994) The development of male and female regenerants by *in vitro* androgenesis in dioecious plant *Melandrium album*. *Ann Bot* 73: 455–459.

Van Niggevecht G, Prentice HC (1985) A note on the sex chromosomes of the *Valencian endemic, Silene diclinis (Caryophyllaceae)*. *Anales Jard Bot Madrid* 41: 267–270.

Vershinin AV, Schwarzacher T, Heslop-Harrison JS (1995) The large-scale genomic organization of repetitive DNA families at the telomeres of rye chromosomes. *Plant Cell* 7: 1823–1833.

Veuskens J, Marie D, Brown SC, Jacobs M, Negrutiu I (1995) Flow sorting of the Y sex chromosome in the dioecious plant *Melandrium album*. *Cytometry* 21: 363–373.

Vyskot B, Araya A, Veuskens J, Negrutiu I, Mouras A (1993) DNA methylation of sex chromosomes in a dioecious plant, *Melandrium album*. *Mol Gen Genet* 239: 219–224.

Westergaard M (1958) The mechanism of sex determination in dioecious flowering plants. *Adv Genet* 9: 217–281.

Winge Ö (1931) X- and Y-linked inheritance in *Melandrium*. *Hereditas* 15: 127–165.

Ye D, Installé P, Ciupercescu D *et al.* (1990) Sex determination in the dioecious *Melandrium*: I. First lessons from androgenic haploids. *Sex Plant Reprod* 3: 179–186.

© 2004 Wiley-Liss, Inc.

Whole Genome Amplification for CGH Analysis: Linker-Adapter PCR as the Method of Choice for Difficult and Limited Samples

Christine Pirker,[1] Maria Raidl,[2] Elisabeth Steiner,[1] Leonilla Elbling,[1] Klaus Holzmann,[1] Sabine Spiegl-Kreinecker,[3] Michaela Aubele,[4] Bettina Grasl-Kraupp,[2] Christine Marosi,[5] Michael Micksche,[1] and Walter Berger[1]*

[1]Division of Applied & Experimental Oncology, Institute of Cancer Research, Vienna University, Vienna, Austria
[2]Division of Toxicology, Institute of Cancer Research, Vienna University, Vienna, Austria
[3]Department of Neurosurgery, Landesnervenklinik Wagner-Jauregg Hospital, Linz, Austria
[4]GSF-National Research Center for Environment and Health, Institute of Pathology, Neuherberg, Germany
[5]Department of Medicine I, University Hospital Vienna, Vienna, Austria

Received 17 November 2003; Revision Received 17 March 2004; Accepted 19 March 2004

Background: Comparative genomic hybridization (CGH) is a powerful method to investigate chromosomal imbalances in tumor cells. However, DNA quantity and quality can be limiting factors for successful CGH analysis. The aim of this study was to investigate the applicability of degenerate oligonucleotide-primed PCR (DOP-PCR) and a recently developed linker-adapter-mediated PCR (LA-PCR) for whole genome amplification for use in CGH, especially for difficult source material.

Methods: We comparatively analyzed DNA of variable quality derived from different cell/tissue types. Additionally, dilution experiments down to the DNA content of a single cell were performed. FISH and/or classical cytogenetic analyses were used as controls.

Results: In the case of high quality DNA samples, both methods were equally suitable for CGH. When analyzing very small amounts of these DNA samples (equivalent to one or a few human diploid cells), DOP-PCR-CGH, but not LA-PCR-CGH, frequently produced false-positive signals (e.g., gains in 1p and 16p, and losses in chromosome 4q). In case of formalin-fixed paraffin-embedded tissues, success rates by LA-PCR-CGH were significantly higher as compared to DOP-PCR-CGH. DNA of minor quality frequently could be analyzed correctly by LA-PCR-CGH, but was prone to give false-positive and/or false-negative results by DOP-PCR-CGH.

Conclusions: LA-PCR is superior to DOP-PCR for amplification of DNA for CGH analysis, especially in the case of very limited or partly degraded source material. © 2004 Wiley-Liss, Inc.

Key terms: comparative genomic hybridization; CGH; degenerate oligonucleotide-primed PCR; DOP-PCR; linker-adapter-mediated PCR; whole genome amplification; paraffin-embedded tissue; degraded DNA

Besides mutations, under- and overrepresentations of tumor-suppressor genes and oncogenes (1), respectively, due to loss or amplification of chromosomal material, might be of decisive importance and a driving force for malignant progression. DNA amplification, especially, is a key mechanism that allows cancer cells to increase expression of critical genes responsible for cell growth and cell cycle progression (2). Comparative genomic hybridization (CGH), first described in 1992 (3), is a powerful technique that enables the detection of imbalanced genomic alterations throughout the genome. CGH experiments require 0.2–2.0 μg of labeled tumor and reference DNA for successful hybridization (4). Therefore, when only small amounts of sample DNA are available, whole genome amplification prior to labeling is required. Especially when CGH is performed in a routine setting, it might be necessary to investigate DNA samples of differing quality and purity, prepared from different institutions, by different methods, from different tissues, which all leads to high demands on reliability and reproducibility of the DNA amplification method.

Contract grant sponsor: Oberösterreichische Krebshilfe; Contract grant sponsor: Jubiläumsfonds der Österreichischen Nationalbank; Contract grant number: 7258.

*Correspondence to: Walter Berger, Institute of Cancer Research, Div. Applied & Experimental Oncology, Vienna University, Borschkegasse 8a, A-1090 Vienna.
E-mail: walter.berger@univie.ac.at

Published online 29 June 2004 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/cyto.a.20060

Degenerate oligonucleotide-primed PCR (DOP-PCR), which was first described in 1992 (4), is currently the most widely used method to amplify whole genome DNA for CGH analysis. DOP-PCR is based on the use of a partially degenerated oligonucleotide primer and PCR amplification under increasingly stringent annealing conditions. Due to technical problems with regard to application for CGH analysis, this method has been frequently modified (5). DNA derived from paraffin-embedded tissues, especially, often causes problems in CGH analysis. This is partly due to DNA degradation per se, but also because the procedures of formalin-fixation and paraffin-embedding might interfere with the PCR reaction. Inhibition of the PCR reaction has also been observed for DNA containing melanin (6,7), a problem that can be overcome by adding bovine serum albumin to the reaction mix. More recently, an optimized strategy for global DNA amplification based on a linker-adapter-mediated PCR method (LA-PCR, also termed SCOMP) has been developed and validated for CGH analysis (8,9). In this method, DNA is first digested by the restriction enzyme MseI, resulting in a fragment size range required for successful CGH. Following digestion, DNA is amplified and labeled via a linker-adapter-PCR, using only a single primer.

In this study, we compared the applicability and quality of DOP-PCR and LA-PCR to amplify whole genome DNA for CGH analyses of difficult, i.e., partly degraded or very limited, DNA samples derived from different sources and prepared by different methods. In summary, we demonstrate that both DOP-PCR and LA-PCR are applicable for DNA amplification with respect to CGH analysis; however, we also demonstrate that the latter method is less prone to give false results in case of very low amounts or minor quality of the DNA samples that are to be analyzed.

MATERIALS AND METHODS

Cell Cultures

Primary cell cultures were established from 48 histopathologically-confirmed human melanomas (10), three glioblastomas (11), and four soft tissue sarcomas. The SK-Mel-28, Calu-6, and HepG2 cell lines were obtained from the American Type Culture Collection (Rockville, MD), and the F2000 embryonic fibroblasts were obtained from Flow, Scotland. All primary cell cultures were established and grown in RPMI 1640 culture medium (Sigma, St. Louis, MO) supplemented with 10% fetal bovine serum and glutamine (Sigma). Cell cultures were frequently checked for *Mycoplasma* contamination.

Paraffin-Embedded and Frozen Material

Samples were obtained from snap-frozen or from formalin-fixed, paraffin-embedded pathology specimens, prepared from resections done at the General Hospital of Vienna and the Landesnervenklinik Wagner-Jauregg Hospital, Linz. Frozen samples were stored in liquid nitrogen. Paraffin-embedded samples were archived at the Departments of Pathology and Dermatology, University Hospital of Vienna, as well as at Wagner-Jauregg Hospital.

Genomic DNA Isolation Procedures

DNA from all cell cultures and blood samples, except that from soft tissue sarcomas, was isolated with the QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions, and stored at –20°C. When comparatively tested against several other kits and methods, in our hands this kit gave the most reliable high-quality DNA extraction results from cultured cells (data not shown). From soft tissue sarcoma primary cell cultures, DNA was extracted using a standard phenol-chloroform-isoamylalcohol method. Briefly, 500 µl lysis buffer (50 mmol Tris, 10 mmol EDTA, 0.5% Na-lauryl-sarcosine) were added to 1.0–1.5 ml of a single-cell suspension. Incubation with 100 µg proteinase K (Sigma) was done at 37–56°C overnight. Then DNA was extracted twice by phenol-chloroform-isoamylalcohol, precipitated by sodium acetate/ethanol, dissolved in aqua bidest (Fresenius Kabi, Graz, Austria), and stored at –20°C. DNA from frozen material was isolated after histopathological examination from ≈50 mg of unstained tissue using the Aqua Pure Genomic Tissue Kit (BioRad, Munich, Germany) according to the manufacturer's instructions. In case of two soft tissue sarcomas, DNA was extracted by phenol-chloroform-isoamylalcohol, as described above. For DNA preparation from paraffin-embedded material of hepatocellular carcinomas and melanomas, three to five sequential sections were mounted onto separate microscope slides. The first one and the last one were stained with hematoxylin and eosin (H/E); the other sections remained unstained. Unstained sections were deparaffinized, and areas of interest were microdissected either with a sterile 20-gauge needle under a microscope or by laser-capture microdissection (PixCell II; Arcturus GmbH, Moerfelden-Walldorf, Germany). In some cases, unstained sections were stained very shortly with hematoxylin for better identification of tumor cells. DNA was then extracted using the Aqua Pure Genomic Tissue Kit following the manufacturer's instructions. In some cases, digestion with proteinase K was prolonged up to three days. For melanomas, the proteinase K digest was done in aqua bidest, and after enzyme inactivation by heat treatment, the sample was directly used for MseI digestion and PCR. For breast carcinomas, DNA was prepared as follows: one section (10 µm) was deparaffinized, the pellet dissolved in Tris-buffer (50 mmol; pH 8.3), digested by proteinase K at 50°C overnight, and the reaction stopped at 95°C for 10 min.

Amplification of DNA by DOP-PCR and LA-PCR

DOP-PCR was used for amplification and labeling of DNA samples and normal reference DNA (Promega, Mannheim, Germany), as described previously (10). In brief, for amplification, 10–100 ng of both tumor and reference DNA was used in a 25 µl reaction mixture: 250 µmol each dATP, dCTP, dGTP, and dTTP, 1% W1-detergent (Sigma, St. Louis, MO), 2 µmol degenerate primer 5′-CCGACTCGAGNNNNNNATGTGG-3′ (VBC, Vienna, Austria), and 0.1

U/μl Super Taq DNA Polymerase (ViennaLab, Vienna, Austria; HT Biotechnology, UK) in Buffer D (Invitrogen, Groningen, The Netherlands). Cycling conditions were as published (10). For the labeling PCR, 1–2 μl of the DOP-PCR product was used for labeling with Dig-11-dUTP (tumor DNA) or Bio-16-dUTP (reference DNA) (Roche, Mannheim, Germany). Labeling was carried out in a 50 μl reaction mixture: 200 μmol each dATP, dCTP, and dGTP, 160 μmol dTTP, 40 μmol Dig-11-dUTP or Bio-16-dUTP, 1% W1-detergent, 2 μmol degenerate primer, and 0.1 U/μl Super Taq DNA Polymerase in TAPS buffer (25 mmol TAPS, 50 mmol KCl, 2 mmol $MgCl_2$, 2.88 μmol β-mercaptoethanol). Labeling PCR conditions were as published (10).

LA-PCR and CGH analysis was originally described by Klein et al. (8), and later termed "SCOMP," for single cell CGH (8,9). In brief, tumor or reference DNA was MseI-digested for 3 h at 37°C in a volume of 5 μl of buffer containing 0.5 μl of 10× One-Phor-All buffer (Amersham, Freiburg, Germany) and 10 units of MseI-enzyme (New England Biolabs, Beverly, MA). Following that, the adapter MseLig12 (5′-TAACTAGCATGC-3′; Biometra, Göttingen, Germany) was ligated to the 5′-TA overhangs, thus providing the binding site for the primer MseLig21 (5′-ATGGGGATTCCGCATGCTAGT-3′). Amplification and labeling PCR were performed exactly as described (10), except that 2 μl of the primary PCR product was used for the labeling PCR. Finally, the linker and the adapter were removed by a second MseI digest and subsequent purification with Micro-bio spin 30 columns (BioRad). All DOP-PCR and LA-PCR products were analyzed by electrophoretic separation on 1% agarose gels.

Comparative Genomic Hybridization (CGH) and Image Analysis

CGH was performed as recently described (10). Increases and decreases in DNA sequence copy numbers were defined by tumor to reference ratios of >1.2 and <0.8. These reference values were established before the study by CGH analysis of normal DNA samples from different sources as specificity control, as well as different mixtures of male and female normal DNA and analysis of X-chromosome material as sensitivity control. Chromosomes 19, 22, and 1p32-pter (3) have not been included in the analyses. To exclude DNA contamination, control PCR reactions were performed without template DNA as a negative control.

Preparation of Metaphase Spreads, FISH, and CDD-Banding

Tumor and normal human metaphase chromosomes were prepared according to standard protocols and FISH analyses with whole chromosome and single locus paints were performed as described (10). BACs for selected gene loci were searched from ENSEMBL (http://www.ensembl.org/) and obtained from the Sanger Center, Cambridge, UK. Chromomycin/Distamycin/DAPI-staining of metaphase chromosomes (CDD-Banding) allowed the simultaneous production of reverse bands and DAPI-bands (12).

Table 1
Success Rates Achieved by DOP-PCR-CGH and LA-PCR-CGH

Sample source	DOP-PCR-CGH experiments, successful/total (%)	LA-PCR-CGH experiments, successful/total (%)
Cell cultures		
Melanoma	54/54 (100)	15/15 (100)
Soft tissue sarcoma	4/4 (100)	4/4 (100)
Glioblastoma	3/3 (100)	3/3 (100)
Hepatocellular carcinoma	1/1 (100)	1/1 (100)
NSCLC	1/1 (100)	1/1 (100)
Embryonic fibroblasts	1/1 (100)	1/1 (100)
Frozen tissues		
Liver	19/19 (100)	3/3 (100)
Soft tissue sarcoma	2/2 (100)	2/2 (100)
Paraffin-embedded tissues		
Liver	8/48 (17)	16/36 (44)
Melanoma	4/12 (33)	21/24 (87.5)
Breast carcinoma	4/4 (100)	4/4 (100)

RESULTS

CGH Success Rates Using DOP-PCR and LA-PCR

CGH experiments that resulted in an adequate hybridization signal of the amplified and labeled DNA samples compatible with the standard settings of the evaluation software (Leica, QFISH) were defined as "successful CGH experiments." Success rates for DOP-PCR- and LA-PCR-CGH experiments for DNA samples derived from living cells, frozen tissues, and paraffin-embedded tissues are summarized in Table 1. All cell culture- and frozen-tissue-derived DNA samples hybridized well to the metaphase spreads and resulted in successful CGH experiments, irrespective of tissue type, DNA extraction method (compare Material and Methods), and amplification by DOP-PCR or LA-PCR. Correspondingly, in all cases DNA was proven to display no signs of degradation in gel electrophoresis. Amplification and labeling by both methods resulted in an adequate range of DNA fragment lengths for CGH (Fig. 1). DOP-PCR-CGH experiments, as well as LA-PCR-CGH experiments, displayed high hybridization intensities and excellent signal-to-noise ratios. In general, LA-PCR-CGH experiments more often showed a smoother hybridization with very low granularity, as compared to DOP-PCR-amplified samples (example shown in Fig. 2). In summary, the data indicate equal quality of the two amplification methods for high quality DNA derived from living cells and frozen sections.

However, differences between the two methods tested became obvious when analyzing DNA from paraffin-embedded tumor tissues. In case of liver samples (n = 80) obtained retrospectively from routinely prepared paraffin blocks, only eight of 48 microdissected samples (17%) resulted in DOP-PCR products that were also successfully applied to CGH experiments. In all other samples, the DOP-PCR product was undetectable or the fragment size was too small to be suitable for CGH (Fig. 1), corresponding to intense DNA degradation in the extract. On the contrary, LA-PCR products of most of these degraded DNA samples showed the same fragment size as seen in case of

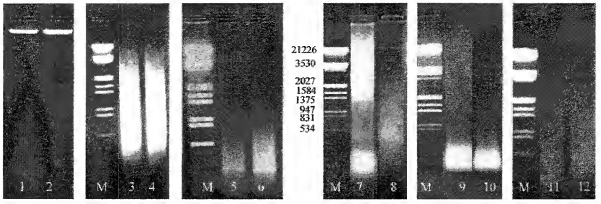

FIG. 1. Impact of DNA quality on DOP-PCR amplification and LA-PCR amplification. Two high quality genomic DNA samples (from frozen tissues of human liver cancer; lanes 1 and 2) and two partly degraded samples (from formalin-fixed, paraffin-embedded tissues of a melanoma and a liver cancer; lane 7 and 8, respectively) are opposed to the respective DOP-PCR amplification products (lanes 3 and 4, as well as 9 and 10, respectively), and LA-PCR amplification products (lane 5 and 6, as well as 11 and 12, respectively). DNA samples and PCR products were separated by 1% agarose gel electrophoresis and stained by ethidium bromide. Notably, the fragment length of the LA-PCR amplification products derived from both degraded and high quality DNA lies within the same range. Molecular weight markers (M) are given in base pairs.

high quality samples (Fig. 1), and 16 of 36 (44%) LA-PCR-CGH experiments were successful with regard to the guidelines for hybridization intensity, signal-to-noise ratio, low background staining, and/or granularity. In case of melanoma tissue sections (n = 26), out of 12 paraffin-embedded tissues, only four (33%, all derived from one pathology department) could be applied to successful DOP-PCR-CGH analysis. In the other eight samples (all from another pathology department), the extracted DNA was too degraded or inaccessible for successful amplification by DOP-PCR. On the contrary, when applying LA-PCR, 21 of 24 experiments (87.5%) led to successful CGH and even allowed comparison of several tumor regions with reliable results (data not shown). In the remaining three LA-PCR-CGH experiments, hybridization signal intensities were too low and background noise too high to score the CGH experiment "successful."

Comparison of DOP-PCR-CGH and LA-PCR-CGH of the Same DNA Sample

A total of 36 DNA samples was comparably analyzed by both methods. In the case of high quality DNA from cell cultures and frozen material prepared by the kit method, both DOP-PCR-CGH and LA-PCR-CGH experiments showed comparable quality and intensity of hybridization. Both methods resulted in identical CGH profiles in all cases analyzed, independent of tumor type and origin (from cell cultures or frozen tissues). All changes checked by FISH analyses with whole chromosome and locus-specific probes were verified (example in Fig. 2). Surprisingly, DOP-PCR-CGH and LA-PCR-CGH led to inconsistent results for four soft tissue sarcomas extracted by a phenol-chloroform-isoamylalcohol method, despite a high molecular size of all DNA samples and a comparable hybridization quality and intensity with both methods. When compared to LA-PCR-CGH and CDD-banding, DOP-PCR-CGH resulted in multiple false-positive signals, whereas false-negative signals were almost absent (Fig. 3).

The tendency toward false results from DOP-PCR-amplified probes, despite CGH experiments scored as successful, was also detected when analyzing a variety of samples from paraffin-embedded tissues. The DNA extracted from several of these samples showed strong degradation (Fig. 1). Hybridization intensity of DOP-PCR-CGH and LA-PCR-CGH was comparable in most of these cases, however, the hybridization quality of DOP-PCR-CGH showed a rather high degree of granularity as well as high background noise when compared to LA-PCR-CGH. For 11 DNA samples derived from paraffin-embedded tissues that displayed distinct signs of DNA degradation, the false results derived from DOP-PCR-CGH, which were avoided by LA-PCR-CGH, mainly concerned chromosomes 4 (10/11; 91%), 17 (10/11; 91%), 13 (7/11; 64%), 1p (6/11; 54%), 5 (5/11; 45%), 9p (5/11; 45%), 12 (4/11; 36%), and 16p (4/11; 36%).

Aged DNA Samples

Two DNA samples derived from the blood of healthy donors and 1 DNA sample derived from a melanoma cell culture were analyzed. All samples had been stored for 3.2 years at −20°C, and had been repeatedly thawed and refrozen during this time. Gel electrophoresis showed in all cases some high molecular weight DNA and a smear of degraded fragments (not shown). LA-PCR-CGH correctly detected the normal karyotypes of the two blood-derived DNA samples of healthy donors (Fig. 4). The aged DNA derived from the melanoma cell culture showed identical results to those obtained by the first CGH analysis three years before. In contrast, DOP-PCR-CGH of all tested DNA samples revealed a tendency towards false gains and losses. Affected loci were almost equal to the ones prone

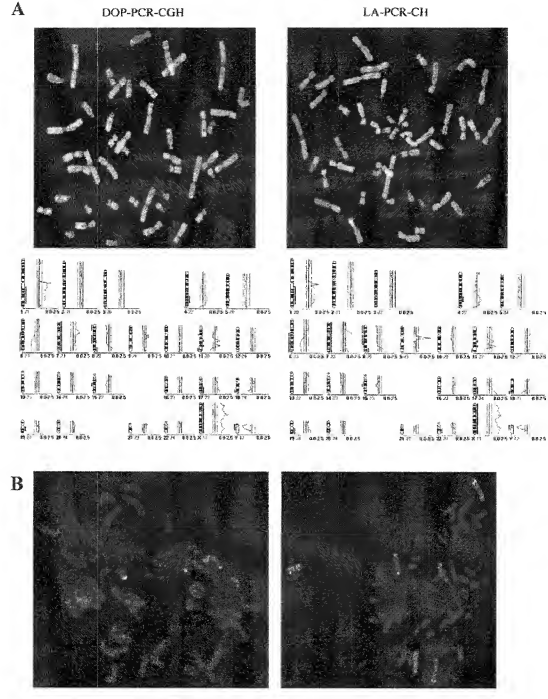

FIG. 2. Comparison of CGH results using a high quality DNA sample. DNA was extracted from a melanoma brain metastasis cell culture. A: Analysis was done by DOP-PCR-CGH and LA-PCR-CGH as indicated. To illustrate hybridization quality, a representative metaphase for DOP-PCR-CGH (left) and LA-PCR-CGH (right) is shown. Mean CGH profiles were calculated from 12 metaphases each. B: FISH analysis of selected tumor metaphase chromosomes was used as control. Left: Chromosome 8 (red) with the gene locus for c-myc on 8q24 (green). Right: Chromosome 1 (green) with the gene locus for "glioma amplified on chromosome 1 protein" (GAC1) on 1q32.1 (yellow).

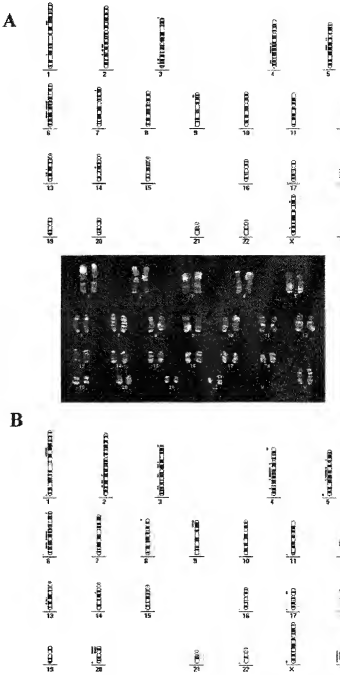

FIG. 3. Comparison of CGH results using DNA samples derived from two soft tissue sarcomas. A: In this case, CGH results (schematic illustration in the upper panel; gains are indicated by gray bars on the right of respective chromosomes, losses as black bars on the left) from DOP-PCR-CGH (inner bars) indicated a series of genomic changes (mainly losses), while LA-PCR-CGH (outer bars) only detected changes in the sex chromosome ratio. CDD-banding of the corresponding tumor metaphase chromosomes (lower panel) showed a normal female karyotype and thus confirmed the results obtained by LA-PCR-CGH. B: In this case, in addition to gains/losses detected by both methods, copy number changes by DOP-PCR-CGH (inner bars) were seen, e.g., on chromosomes 1p, 2q, 4q, 5q, 6q, 9p, and 13q, as well as on chromosome 19. CDD-banding of the corresponding tumor metaphase chromosomes revealed a highly complex karyotype and widely confirmed the results obtained by LA-PCR-CGH (data not shown).

to give false results in the degraded paraffin-embedded samples (Fig. 4).

Dilution Experiments

In order to test the efficiency of DOP-PCR and LA-PCR to amplify very small amounts of DNA corresponding to one single diploid cell (5–7 pg) and consequently produce reliable CGH results, we performed DNA dilution experiments. DNA amounts analyzed ranged from 500 ng to 2 pg. PCR products derived from DNA samples containing 20 pg or more, produced in both methods reliable and identical CGH profiles (Fig. 5A and B). In case of less than 20 pg of template DNA, DOP-PCR-CGH showed additional to sample-specific changes a series of false-positive and

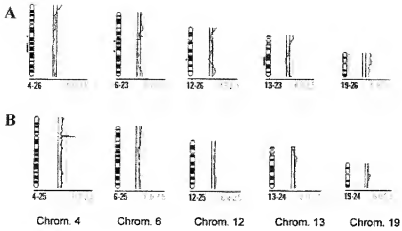

FIG. 4. Comparison of CGH results from an aged DNA sample. DNA was derived from blood of a healthy donor stored for 3.2 years at -20°C, and repeatedly thawed and refrozen during this time. CGH profiles for selected chromosomes as indicated are shown. A: DOP-PCR-CGH revealed several gains and losses that would have been avoided by redefining thresholds to >1.25 and <0.75 (not shown). B: LA-PCR-CGH correctly indicated the normal karyotype at the standard threshold >1.2 and <0.8.

false-negative results, as well as changes in the green to red ratio of certain chromosomal regions (Fig. 5C). These were mainly characterized by a decrease of gained regions and an increase of lost regions, as well as the excessive amplification of heterochromatic regions, e.g. at the short arms of acrocentric chromosomes. Also, those chromosomal regions described as critical in CGH (13) (1p32-pter, 16p, 19, and 22) displayed increased gains as compared to the original experiment. On the other hand, all LA-PCR-CGH experiments using only ≈5 pg DNA showed reliable results, with profiles identical to the profiles obtained from undiluted DNA samples (Fig. 5B and D). LA-PCR-CGH experiments from less than 5 pg of template DNA could not be evaluated, because the hybridization intensity was too weak (not shown).

DISCUSSION

CGH is a helpful screening tool to detect DNA sequence copy number imbalances in malignant cells (3,14,15). However, investigations of early lesions, specific tumor regions, and small metastatic lesions led to the necessity to perform CGH from minute amounts of DNA. In order to solve this problem, several methods for whole genome DNA amplification have been developed. Due to its simplicity and success rate, DOP-PCR has become the method of choice (3). However, technical problems in case of DNA prepared from paraffin-embedded samples led to several re-evaluations of this technique (4–7). Recently, another method, LA-PCR, was introduced, suggesting that it would allow whole genome amplification down to the level of one metastatic cell (16,17). Our study aimed to investigate to what extent these two methods would be suitable for DNA sample preparation in a routine clinical setting, which might include inefficiently prepared paraffin-tissues, aged samples, and DNA prepared by completely different methods. Our data demonstrate that DOP-PCR as well as LA-PCR are highly efficient tools for whole genome amplification in order to perform CGH from target tissues down to the level of a few cells. However, especially in the case of partly degraded DNA extracted from archived, paraffin-embedded tissues, LA-PCR-CGH provided a significantly higher success rate than DOP-PCR-CGH. Moreover, in partly degraded, but also in aged DNA samples, in DNA samples extracted by certain precipitation methods, and in the case of amplification of very low amounts of DNA (equivalent to one to five diploid cells), LA-PCR-CGH delivered distinctly more reliable results than DOP-PCR-CGH. In summary, our data demonstrate, in accordance with a previous report (8), that LA-PCR is of superior quality and reliability as compared to DOP-PCR for routine amplification of DNA from difficult and limited sources in order to perform CGH.

The observed differences between the two methods with respect to both the success rates and the reliability of the CGH results are thought to be based on the principal differences between the two amplification methods. DOP-PCR uses a multitude (4^6) of different partially degenerated primers that anneal under increasingly stringent PCR conditions at binding sites that are supposed to be present about every 4 kb of genomic DNA (4). On the contrary, LA-PCR is preceded by a DNA-digestion-step by MseI that allows, via ligation of an adapter, the binding of one single primer, and the subsequent amplification of the whole genome using only one primer. The exactly defined MseI-restriction sites are randomly present about every 200–300 bp (8), which makes this method more suitable for the amplification of already smaller DNA fragments.

We focused on the analysis of formalin-fixed, paraffin-embedded tissues of two tumor entities, namely malignant melanoma and hepatocellular carcinoma, which are known to be critical for whole genome amplification. For primary melanomas, microdissectable parts are frequently very small because of small tumor mass due to tumor detection at early stages. Additionally, high amounts of melanin can hamper the PCR procedure (7). Despite these limitations, and despite the use of aged (up to eight years) melanoma paraffin blocks, LA-PCR raised the CGH success rate to almost 90% of melanoma samples, while only a limited proportion of these specimens (33%) could be analyzed using DOP-PCR. This success rate resembles the one published previously for

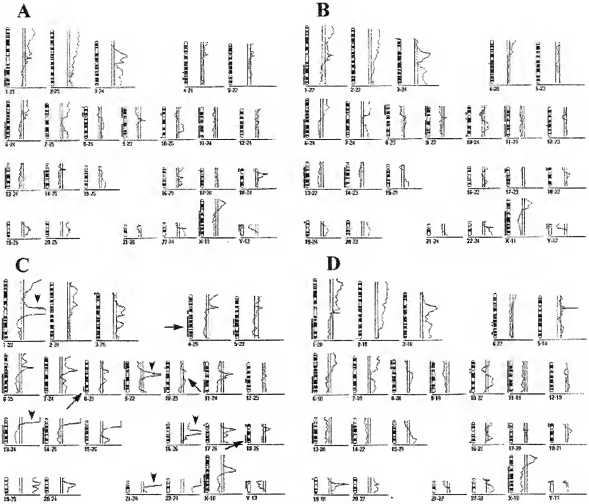

FIG. 5. Dilution experiments using DNA from a melanoma brain metastasis cell culture. The amount of DNA prior to amplification was 500 ng in the undiluted samples (A,B) and 5 pg in the diluted samples (C,D). DOP-PCR-CGH (A,C) and LA-PCR-CGH profiles (B,C) of the undiluted sample and the diluted sample are shown. Whereas LA-PCR-CGH of the diluted sample showed the identical results as obtained by both CGH experiments of the undiluted sample, DOP-PCR-CGH of the diluted sample (C) showed additional changes (examples indicated by arrows) as well as excessive gains at heterochromatic regions (arrowheads).

LA-PCR-CGH with respect to tumors of mixed origin (9). Also for formalin-fixed, paraffin-embedded liver samples, LA-PCR was clearly superior to DOP-PCR. However, as compared to the data above, success rates (44% versus 17% for LA-PCR and DOP-PCR, respectively) were relatively low. This might have been caused by the high amounts of degrading enzymes in the liver as well as enhanced DNA-degradation by formalin fixation (18,19). Stoecklein et al. (9) suggested the sample age of formalin-fixed, paraffin-embedded tissues as a decisive factor limiting successful amplification, especially by DOP-PCR. In contrast, we did not find any correlation between sample age and degradation of the extracted DNA in human liver samples. Our data, rather, suggest proper handling of the surgery specimen as a key requisite for DNA-amplification and CGH analysis. It is of utmost importance that liver samples are fixed immediately after surgery and that the fixation time is long enough, depending on the sample size (20). This is additionally corroborated by the fact that we monitored distinct differences in success rates between samples obtained from different pathology departments. In accordance with our findings, Dietmaier et al. (19) suggested that successful whole genome amplification for mutation analysis is critically dependent on a very quick sample procession (<30 min), and problematic when analyzing paraffin-embedded tissue samples. Thus, the use of LA-PCR is especially recommended in situations where CGH has to be routinely performed from samples derived from different tissues or from institutions with different fixation and embedding techniques.

Unexpectedly, comparable problems to those found with paraffin-embedded tissues were observed for DNA preparations from cell cultures using a precipitation

method in combination with analysis by DOP-PCR-CGH. Despite the presence of high molecular weight DNA without signs of protein contamination, DOP-PCR-CGH tended to give false-positive results that were absent in the case of LA-PCR-CGH. The observed problems suggest that incomplete precipitation and/or impaired solubility of DNA from specific chromosomal regions might be limiting for DNA amplification and CGH analysis. The resulting DNA aggregates might be inaccessible for DOP-PCR, causing false-positive results in several chromosomal regions. LA-PCR is preceded by an enzyme digestion-step and a following ligation reaction, which might both have a positive influence on the accessibility of the DNA.

Both, DOP-PCR and LA-PCR (also termed SCOMP) have been suggested to be suitable for analysis of very minor amounts of DNA, equivalent to one or a few cells (9,16,17,21–23). We compared the sensitivity and efficiency of both amplification procedures in a series of DNA dilution experiments. For either amplification method, DNA amounts from 300 ng to 20 pg resulted in reliable experiments, with identical results obtained from undiluted and diluted DNA samples. These data are in good agreement with previous studies in which the minimal DNA amounts for successful DOP-PCR-amplification and CGH were described to range between 12.5–50 pg (5,21,23). At lower amounts of target DNA, equivalent to one or a few diploid cells, we found LA-PCR to be clearly superior as compared to DOP-PCR. LA-PCR-CGH experiments using DNA amounts of only 5 pg for amplification showed good hybridization quality, as well as the identical profiles obtained by the corresponding undiluted sample. On the contrary, DOP-PCR-CGH experiments using less than 20 pg of template DNA showed a tendency to false results due to loss of specifically gained regions and increased losses of already underrepresented regions, as well as gains of heterochromatic regions and of chromosomal regions already described as critical in CGH (13). These results are in accordance with several studies demonstrating the applicability of LA-PCR for DNA analysis from one or a few cells (9), allowing detection of the genetic heterogeneity (16) and combined characterization of genetic alterations and gene expression (17).

In summary, our results demonstrate that LA-PCR leads to a higher number of successful CGH experiments and more reliable results when compared with DOP-PCR, especially if critical DNA samples derived from formalin-fixed, paraffin-embedded tissues, as well as very low amounts of DNA have to be investigated.

ACKNOWLEDGMENTS

We thank B. Stögermayer, M. Eisenbauer, and V. Bachinger for excellent technical assistance.

LITERATURE CITED

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100: 57–70.
2. Knuutila S, Bjorkqvist AM, Autio K, Tarkkanen M, Wolf M, Monni O, Szymanska J, Larramendy ML, Tapper J, Pere H, El-Rifai W, Hemmer S, Wasenius VM, Vidgren V, Zhu Y. DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. Am J Pathol 1998;152:1107–1123.
3. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992;258:818–821.
4. Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 1992; 13:718–725.
5. Speicher MR, du Manoir S, Schrock E, Holtgreve-Grez H, Schoell B, Lengauer C, Cremer T, Ried T. Molecular cytogenetic analysis of formalin-fixed, paraffin-embedded solid tumors by comparative genomic hybridization after universal DNA-amplification. Hum Mol Genet 1993;2:1907–1914.
6. Price K, Linge C. The presence of melanin in genomic DNA isolated from pigmented cell lines interferes with successful polymerase chain reaction: a solution. Melanoma Res 1999;9:5–9.
7. Eckhart L, Bach J, Ban J, Tschachler E. Melanin binds reversibly to thermostable DNA polymerase and inhibits its activity. Biochem Biophys Res Commun 2000;271:726–730.
8. Klein CA, Schmidt-Kittler O, Schardt JA, Pantel K, Speicher MR, Riethmuller G. Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proc Natl Acad Sci USA 1999;96:4494–4499.
9. Stoecklein NH, Erbersdobler A, Schmidt-Kittler O, Diebold J, Schardt JA, Izbicki JR, Klein CA. SCOMP is superior to degenerated oligonucleotide primed-polymerase chain reaction for global amplification of minute amounts of DNA from microdissected archival tissue samples. Am J Pathol 2002;161:43–51.
10. Pirker C, Holzmann K, Spiegl-Kreinecker S, Elbling L, Thallinger C, Pehamberger H, Micksche M, Berger W. Chromosomal imbalances in primary and metastatic melanomas: over-representation of essential telomerase genes. Melanoma Res 2003;13:483–492.
11. Berger W, Spiegl-Kreinecker S, Buchroithner J, Elbling L, Pirker C, Fischer J, Micksche M. Overexpression of the human major vault protein in astrocytic brain tumor cells. Int J Cancer 2001;94:377–382.
12. Schweizer D, Ambros PF. Chromosome banding. In: Gosden JR, editor. Methods in molecular biology, Vol. 29. Totowa, NJ: Humana Press Inc.; 1994. p 98–111.
13. Kallioniemi OP, Kallioniemi A, Piper J, Isola J, Waldman FM, Gray JW, Pinkel D. Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors. Genes Chromosomes Cancer 1994;10:231–243.
14. Nacheva EP, Grace CD, Bittner M, Ledbetter DH, Jenkins RB, Green AR. Comparative genomic hybridization: a comparison with molecular and cytogenetic analysis. Cancer Genet Cytogenet 1998;100:93–105.
15. Rooney PH, Murray GI, Stevenson DA, Haites NE, Cassidy J, McLeod HL. Comparative genomic hybridization and chromosomal instability in solid tumours. Br J Cancer 1999;80:862–873.
16. Klein CA, Blankenstein TJ, Schmidt-Kittler O, Petronio M, Polzer B, Stoecklein NH, Riethmuller G. Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. Lancet 2002;360:683–689.
17. Klein CA, Seidl S, Petat-Dutter K, Offner S, Geigl JB, Schmidt-Kittler O, Wendler N, Passlick B, Huber RM, Schlimok G, Baeuerle PA, Riethmuller G. Combined transcriptome and genome analysis of single micrometastatic cells. Nat Biotechnol 2002;20:387–392.
18. Wiegand P, Dombover J, Brinkmann B. DNA degradation in formalin fixed tissues. Pathologe 1996;17:451–454.
19. Dietmaier W, Hartmann A, Wallinger S, Heinmoller E, Kerner T, Endl E, Jauch KW, Hofstadter F, Ruschoff J. Multiple mutation analyses in single tumor cells with improved whole genome amplification. Am J Pathol 1999;154:83–95.
20. Yagi N, Satonaka K, Horio M, Shimogaki H, Tokuda Y, Maeda S. The role of DNase and EDTA on DNA degradation in formaldehyde fixed tissues. Biotech Histochem 1996;71:123–129.
21. Huang Q, Schantz SP, Rao PH, Mo J, McCormick SA, Chaganti RS. Improving degenerate oligonucleotide primed PCR-comparative genomic hybridization for analysis of DNA copy number changes in tumors. Genes Chromosomes Cancer 2000;28:395–403.
22. Hirose Y, Aldape K, Takahashi M, Berger MS, Feuerstein BG. Tissue microdissection and degenerate oligonucleotide primed-polymerase chain reaction (DOP-PCR) is an effective method to analyze genetic aberrations in invasive tumors. J Mol Diagn 2001;3:62–67.
23. Cheung VG, Nelson SF. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. Proc Natl Acad Sci USA 1996;93:14676–14679.

BMC Genomics

BioMed Central

Research article

Open Access

Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing

Robert Pinard[1], Alex de Winter[1], Gary J Sarkis[1], Mark B Gerstein[2], Karrie R Tartaro[1], Ramona N Plant[1], Michael Egholm[1], Jonathan M Rothberg[1] and John H Leamon*[1]

Address: [1]454 Life Sciences, 20 Commercial Street, Branford CT 06405, USA and [2]MB&B Department, Yale University, 266 Whitney Ave., New Haven CT 06520, USA

Email: Robert Pinard - robert.pinard@historx.com; Alex de Winter - adewinter@pacificbiosciences.com; Gary J Sarkis - gary.sarkis@ikonisys.com; Mark B Gerstein - mark.gerstein@yale.edu; Karrie R Tartaro - ktartaro@454.com; Ramona N Plant - ramona.2.plant@gsk.com; Michael Egholm - megholm@454.com; Jonathan M Rothberg - jrothberg@454.com; John H Leamon* - jleamon@454.com

* Corresponding author

Published: 23 August 2006

BMC Genomics 2006, 7:216 doi:10.1186/1471-2164-7-216

Received: 24 May 2006

Accepted: 23 August 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/216

© 2006 Pinard et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Whole genome amplification is an increasingly common technique through which minute amounts of DNA can be multiplied to generate quantities suitable for genetic testing and analysis. Questions of amplification-induced error and template bias generated by these methods have previously been addressed through either small scale (SNPs) or large scale (CGH array, FISH) methodologies. Here we utilized whole genome sequencing to assess amplification-induced bias in both coding and non-coding regions of two bacterial genomes. *Halobacterium* species NRC-1 DNA and *Campylobacter jejuni* were amplified by several common, commercially available protocols: multiple displacement amplification, primer extension pre-amplification and degenerate oligonucleotide primed PCR. The amplification-induced bias of each method was assessed by sequencing both genomes in their entirety using the 454 Sequencing System technology and comparing the results with those obtained from unamplified controls.

Results: All amplification methodologies induced statistically significant bias relative to the unamplified control. For the *Halobacterium* species NRC-1 genome, assessed at 100 base resolution, the D-statistics from GenomiPhi-amplified material were 119 times greater than those from unamplified material, 164.0 times greater for Repli-G, 165.0 times greater for PEP-PCR and 252.0 times greater than the unamplified controls for DOP-PCR. For *Campylobacter jejuni*, also analyzed at 100 base resolution, the D-statistics from GenomiPhi-amplified material were 15 times greater than those from unamplified material, 19.8 times greater for Repli-G, 61.8 times greater for PEP-PCR and 220.5 times greater than the unamplified controls for DOP-PCR.

Conclusion: Of the amplification methodologies examined in this paper, the multiple displacement amplification products generated the least bias, and produced significantly higher yields of amplified DNA.

Background

Continued improvement in sequencing quality, combined with increasingly sophisticated bioinformatic analysis of sequence data, has increased the relevance of whole genome sequencing to many fields of biological science including the pharmaceutical industry, agriculture, national defence and medicine [1]. Similarly, the increased availability of sequence data has served to support increasingly complex and informative comparative genomic studies. In both cases, the enhanced relevance and power of genomic comparisons have, in turn, furthered demand for still more sequence data, with the goal of comparing entire genomes. Although high-throughput sequencing methodologies have been developed to accommodate the demand for sequence output, they consume large amounts of a valuable input: genomic DNA. For example, a Taq-man based whole genome association study of 300,000 SNPs would require approximately 9 mg of genomic DNA, more than obtained in routine clinical blood samples [2]. Less input DNA is required for a genome-wide, microarray-based survey restricted to known mutations, but even this would require some form of amplification step [3].

Other applications also place a high demand on potentially scarce DNA. Emerging relationships between specific genotypes and risk factors or disease states have focused attention on DNA samples that are of great medical/scientific importance, but of limited supply, such as tumor samples, lavages, buccal swabs, or samples generated by laser capture microscopy [4]. While laser capture offers single cell accuracy and both lavages and swabs permit minimal patient invasion and discomfort, these methodologies produce far less genomic DNA than less precise, more invasive techniques [5-7]. Some inherently rare samples, such as difficult to culture micro-organisms [8] or genes from an individual bacterium [9] are of great scientific interest, but cannot be sequenced by current technologies without pre-amplification [10,11]. Considerable interest also exists in sequencing low abundance DNA from museum or fossil specimens, although amplification of these samples must address issues of degradation [12] and contamination [13,14]. High rates of consumption, combined with high demand from the scientific community, may result in hard decisions restricting access to these limited or irreplaceable samples.

Whole genome amplification (WGA) can potentially eliminate DNA as a limiting factor for genetic assays. However, in order to fulfil this role, WGA must satisfy some basic requirements. First, the amplification process should be highly accurate, so as to avoid introducing an undue number of errors. Second, amplification should not induce a bias in the distribution of the product DNA. Third, a high amplification factor is required, so that the WGA generates a useful amount of DNA from small starting samples. Finally, the WGA method should be applicable to a wide array of genomes. For maximal efficiency, the WGA protocol would be universally applicable, without need for separate optimization for each sample. In this paper we will address the latter three points – bias, yield, and applicability to two different genomes – leaving the more complicated studies of both amplification fidelity and the sequence-specific causes of bias for another time.

Three primary forms of WGA have been developed: multiple displacement amplification (MDA) [15,16], primer extension preamplification (PEP) [17], and degenerate oligonucleotide primed PCR (DOP) [18]. These WGA methods have been compared in previous papers, but these comparisons have been limited in scale. The authors either scanned individual nucleotide mutations for SNP analysis, or used comparative genomic hybridization (CGH) or fluorescence in situ hybridization (FISH) to scan large regions of the genome [15,19-22]. In the SNP analyses, the comparison is at a high resolution, but small in scope, while for CGH and FISH, the comparison is at a low resolution (since these methods are extremely forgiving of point mutation errors), but large in scope. The different methodology resolutions can therefore report differing levels of bias, as described in a recent study of φ29 fidelity using both direct sequencing and array hybridization of 10000 SNPs [23]. While array-hybridization results revealed whole genome amplification-related loss of 6 regions (approximately 5.56 Mb) and under-representation of another 8 regions, SNP calls from amplified DNA were not statistically different from those of unamplified material [23]. Ideally, any comparison of WGA methods would investigate amplification bias across entire genomes at the highest resolution possible.

For this paper, we have used MDA, PEP and DOP to amplify two bacterial genomes, *Halobacterium* species NRC-1 with a relatively high 66% GC content (derived from the 68% GC main chromosome and two, lower GC content associated minichromosomes, of which at least one, pNRC100, is multicopy [24] – see next section for more detail) and *Campylobacter jejuni* with a single, 31% GC content chromosome. After making libraries of the resultant amplified genomes, we sequenced the libraries using the 454 Sequencing System [25], and enumerated reads initiating within various sized windows of the respective genome, with a maximum resolution of 10 bases across the entire length of the genome. We then conducted sequence-based karyotyping on the amplified and control genomes, and were thus able to generate a high-resolution comparison, encompassing both coding and non-coding regions, of the coverage bias induced by WGA methods across complete bacterial genomes.

Results

Amplification yield

DNA samples were assayed by UV absorption at 260 nm to determine their concentration after amplification. The amount of input DNA was held constant at 25 ng for all methods. Averaged across both genomes, GenomiPhi generated 16.1 μg of DNA, a 640-fold amplification, Repli-G amplified input DNA 2100 fold to 53.6 μg, PEP generated 3.0 μg, a 120-fold increase, and DOP amplified the input DNA 92-fold to 2.3 μg.

Data analysis

For each amplification method, sequencing reads that mapped to the target genome at 95% or greater accuracy were pooled from three or more individual sequencing runs. The percentage of sequences that mapped to the genome varied depending on the amplification method utilized: roughly 60% of the unamplified and MDA amplified reads mapped at 95% accuracy or better, while 40% of the PEP and 20% of the DOP samples mapped to their target genomes. The number of pooled, mapped reads for each of the amplified samples exceeded 150,000, while the number of pooled control reads was approximately 1,500,000 for each genome, reflecting the larger number of samples drawn from these pools to assess variability within the controls.

Analysis populations composed of 100,000 unique reads, their start position (in base pairs), read length (in bases) and their orientation (forward or reverse) on the reference genome were randomly sampled from each of the pooled sequences from amplified genomic material. Five separate analysis populations were generated from each of the control sequence pools to determine the degree of variation within the unamplified reads. Following the generation of the analysis populations, the total genome coverage in bases was determined for each population and both genomes, and the percent of total coverage calculated for each in Table 1.

The GC content of the sequenced reads was determined for each genome and amplification methodology. The FASTA files generated for each of the one hundred thousand reads resulting from each amplification method were analyzed for GC content, and the mean GC content and standard deviation for each method was calculated. Welsh's two sample t-test was employed to compare the mean GC content from each test population against the reference population for each genome. The 95% confi-

Table 1: Comparison of genome coverage. Coverage was derived from the individual sequences generated from either unamplified control or whole genome amplified samples.

Halobacterium species NRC-1			
Sample	**Total Bases Sequenced**	**Nonredundant Bases Sequenced**	**Percent of Genome**
Unamplified Control	9543911	2183656	84.9%
Unamplified Replicate 1	9540329	2188525	85.1%
Unamplified Replicate 2	9540870	2184161	85.0%
Unamplified Replicate 3	9541768	2179647	84.8%
Unamplified Replicate 4	9543517	2183939	84.9%
Averaged Replicates	9541621	2184068	84.9%
GenomiPhi	9736445	1287564	50.1%
Repli-G	9432145	933686	36.3%
PEP-PCR	9410215	921446	35.8%
DOP-PCR	9010086	249571	9.7%
Campylobacter jejuni			
Sample	**Total Bases Sequenced**	**Nonredundant Bases Sequenced**	**Percent of Genome**
Unamplified Control	10605551	1635277	99.6%
Unamplified Replicate 1	10605882	1636507	99.7%
Unamplified Replicate 2	10592192	1636033	99.7%
Unamplified Replicate 3	10603419	1636620	99.7%
Unamplified Replicate 4	10588370	1637090	99.7%
Averaged Replicates	10597466	1636563	99.7%
GenomiPhi	10380921	1623858	98.9%
Repli-G	10939177	1625276	99.0%
PEP-PCR	10313911	1549641	94.4%
DOP-PCR	9605188	278645	17.0%

Table 2: Comparison of *Holobocterium* GC coverage. Mean sequence GC contents, standard deviations, P-values and 95% Confidence limits around differences between the test and unamplified control population means for *Holobocterium species NRC-1*.

Population	P Value	Unamplified Control Mean GC Content	Test Population Mean GC Content	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Unamplified Replicate 1	0.32	61.21 ± 8.6%	61.25 ± 8.6%	-0.11%	0.04%
Unamplified Replicate 2	0.88		61.21 ± 8.6%	-0.07%	0.08%
Unamplified Replicate 3	0.92		61.21 ± 8.6%	-0.07%	0.08%
Unamplified Replicate 4	0.23		61.26 ± 8.6%	-0.12%	0.03%
GenomiPhi	< 0.001		58.29 ± 7.7%	2.85%	2.99%
Repli-G	< 0.001		57.36 ± 7.4%	3.78%	3.92%
PEP-PCR	< 0.001		56.25 ± 7.2%	4.90%	5.04%
DOP-PCR	< 0.001		55.20 ± 7.0%	5.94%	6.08%

dence interval around the difference between the means, and the corresponding P-value were recorded in Tables 2 and 3. The sequencing results for both genomes were also analyzed for the type and size of homopolymers covered in the sequencing and summarized in Figures 1 and 2.

The reference genome was then subdivided into bins of a specific number of bases, and the number of reads that started in each bin was recorded for every population. Additionally, the relationship between coverage depth and the presence of minichromosomes or genomic repeat regions was examined for 100 base genomic bins (See Table 4.). For the purposes of this paper and this sequencing technology, genomic repeats were defined as regions which were 95% identical across 100 bases. The genomic location of the repeat regions relative to the counts per 100 base bins from unamplified Halobacterium and C. jejuni controls is shown in Figure 3A and 3B.

For each sample, the number of reads initiated in each bin was compared to the number found in the same bin of the respective unamplified control. The ratio of the two numbers was computed for each bin, and the maximum (representing over-amplification relative to the control) and minimum ratio (representing under-amplification or sequence loss relative to the control) were recorded in Table 5. As an identical number of reads were used for each sample, the average numbers of reads obtained per bin were identical for each treatment, requiring a more sophisticated statistical assessment to accurately assess potential bias.

The empirical cumulative frequency distribution (ECDF) of the reads per bin for *Halobacterium* and *C. jejuni* are shown in Figures 4A and 4B respectively. The ECDF represents the cumulative distribution of the number of counts per bin, reporting the cumulative proportion of bins with counts equal or less than the value on the X axis. It was expected that the counts per bin would follow a Poisson distribution, and some bins in each sample would doubtlessly contain outliers. To address relative bias we wanted to compare differences between the read distributions obtained for each sample, rather than comparing each sample to a model distribution. As a result, the non-parametric, distribution-free Kolmogorov-Smirnov test (KS-test) with its associated D-statistic was used for subsequent analysis.

The Kolmogorov-Smirnov test (KS-test) and D-statistic

The KS-test was employed to assess the deviation between the number of counts per bin that occurred between two genomes (one test, one reference), using the maximal deviation (D-statistic) between the cumulative count distributions of the populations being compared (for a visual

Table 3: Comparison of *Compylobacter jejuni* GC coverage. Mean sequence GC contents, standard deviations, P-values and 95% Confidence limits around differences between the test and unamplified reference population means for *Compylobocter jejuni*.

Population	P Value	Unamplified Control Mean GC Content	Test Population Mean GC Content	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Unamplified Replicate 1	0.65	31.15 ± 6.7%	31.16 ± 6.7%	-0.07%	0.05%
Unamplified Replicate 2	0.75		31.16 ± 6.7%	-0.07%	0.05%
Unamplified Replicate 3	0.83		31.15 ± 6.7%	-0.07%	0.05%
Unamplified Replicate 4	0.75		31.16 ± 6.7%	-0.07%	0.05%
GenomiPhi	< 0.001		31.36 ± 6.8%	-0.28%	-0.16%
Repli-G	< 0.001		31.60 ± 6.7%	-0.52%	-0.40%
PEP-PCR	< 0.001		33.92 ± 8.1%	-2.84%	-2.71%
DOP-PCR	< 0.001		30.79 ± 5.9%	0.30%	0.41%

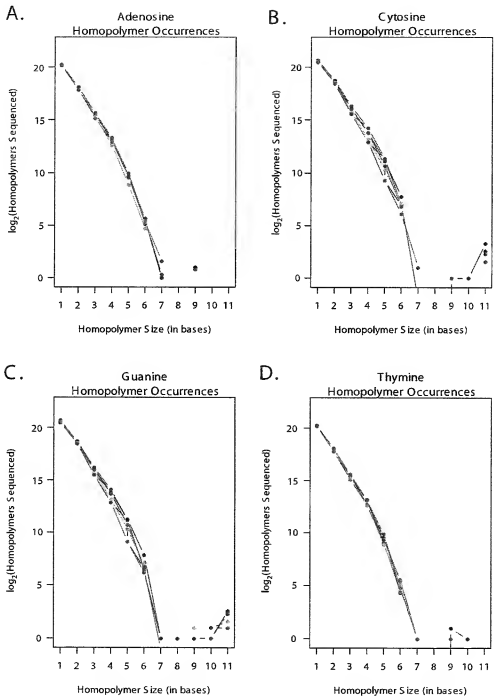

Figure 1

A – D: Hompolymer coverage in *Halobacterium* sequence reads. **A**. A $\log_2$plot illustrating the total number of adenosine homopolymers sequenced in reads generated by control and amplified populations of *Halobacterium* species NRC-1 DNA. The data from the unamplified replicate population are shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B**. As Figure 1A., but for Cytosine homopolymers. **C**. As Figure 1A., but for Guanine homopolymers. **D**. As Figure 1A., but for Thymine homopolymers.

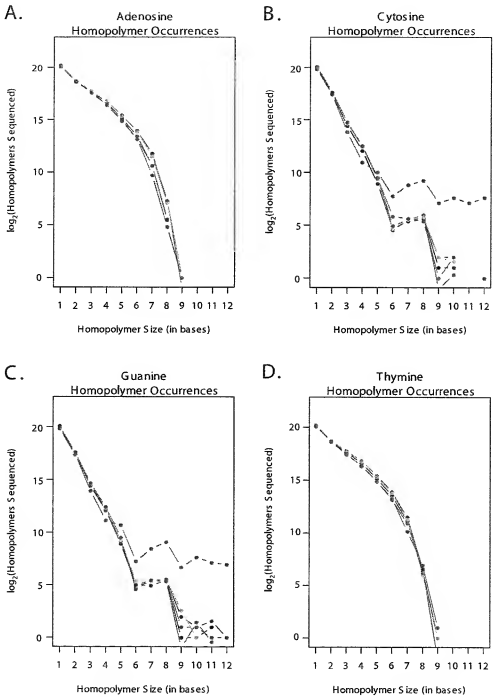

Figure 2

A – D: Hompolymer coverage in *Campylobacter jejuni* sequence reads. A. A $\log_2$ plot illustrating the total number of adenosine homopolymers sequenced in reads generated by control and amplified populations of *Campylobacter jejuni* DNA. The data from the unamplified replicate population are shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B.** As Figure 2A., but for Cytosine homopolymers. **C.** As Figure 2A., but for Guanine homopolymers. **D.** As Figure 2A., but for Thymine homopolymers.

Table 4: Comparison of chromosomal coverage. Sequence coverage per 100 base bin is shown for both repeat and unique regions from unamplified control or whole genome amplified samples.

Halobacterium species NRC-1 main (NRC1) and minichromomes (pNRC100 and pNRC200)

Chromosome	Sample	Reads per unique bin	Reads per repeat bin	Percent chromosomal reads/genome	Percent unique chromosomal reads/genome	Percent repeat chromosomal reads/genome	Percent unique reads/ chromosomal reads	Percent repeat reads/ chromosomal reads
NRC1 2.01 Mb 68% GC 1.4% repeats	Control	1.91	22.38	44%	38%	6%	86%	14%
	Average Unamplified Replicates	1.92	22.18	44%	38%	6%	86%	14%
	GenomiPhi	0.83	30.92	25%	16%	9%	66%	35%
	Repli-G	0.55	36.23	21%	11%	10%	52%	48%
	PEP-PCR	0.69	28.67	22%	14%	8%	63%	37%
	DOP-PCR	0.38	28.07	15%	8%	8%	49%	51%
pNRC100 191Kb 57.9% GC 79.4% repeats	Control	6.42	13.84	24%	3%	21%	11%	89%
	Average Unamplified Replicates	6.51	13.77	23%	3%	21%	11%	89%
	GenomiPhi	9.47	18.64	32%	4%	28%	12%	88%
	Repli-G	9.81	19.68	34%	4%	30%	12%	88%
	PEP-PCR	11.09	20.51	35%	4%	31%	13%	88%
	DOP-PCR	2.59	20.19	32%	1%	31%	3%	97%
pNRC200 365 Kb 59.2% GC 45.0% repeats	Control	3.78	14.97	32%	8%	25%	24%	76%
	Average Unamplified Replicates	3.85	14.89	32%	8%	24%	24%	76%
	GenomiPhi	4.86	20.19	43%	10%	33%	23%	77%
	Repli-G	4.54	21.92	45%	9%	36%	20%	80%
	PEP-PCR	3.76	21.40	43%	8%	35%	18%	82%
	DOP-PCR	7.65	22.92	53%	15%	38%	29%	71%

Campylobacter jejuni single chromosome

Chromosome	Sample	Reads per unique bin	Reads per repeat bin	Percent chromosomal reads/genome	Percent unique chromosomal reads/genome	Percent repeat chromosomal reads/genome	Percent unique reads/ chromosomal reads	Percent repeat reads/ chromosomal reads
C. jejuni 1.64 Mb 31% GC 2.71% repeat	Control	5.90	12.94	N/A	94%	6%	N/A	N/A
	Average Unamplified Replicates	5.90	12.97		94%	6%		
	GenomiPhi	5.96	10.81		95%	5%		
	Repli-G	5.87	13.85		94%	6%		
	PEP-PCR	5.35	32.42		86%	15%		
	DOP-PCR	6.12	4.91		98%	2%		

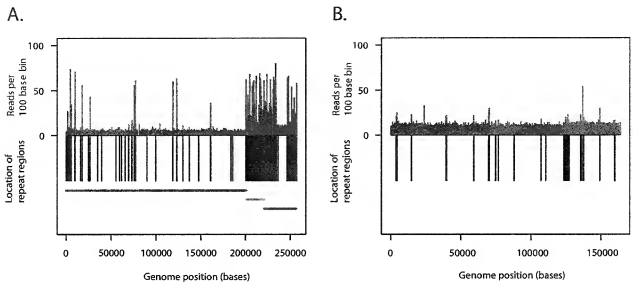

Figure 3

A & B: Comparison of control sequence coverage versus repeat region location. Distribution of sequence coverage from unamplified control *Halobacterium* species NRC-1 population as determined by sequence based karyotyping at 100 base resolution relative to repeat region and chromosome location. **A.** Counts per bin are displayed above the X axis in red, repeat regions are shown below the X axis in black. The X axis displays the length of the genome in bases. The relative location of the three *Halobacterium* chromosomes are shown by the horizontal bars below the X axis, NRC1 is green, pNRC100 is gold, and pNRC200 is purple. **B.** As for Figure 3A, but for *Campylobacter jejuni*. No chromosome bars are included as *C. jejuni* is comprised of a single chromosome.

demonstration of the D-statistic, please see the inset on Figure 4A). The D-statistic for each comparison reported the maximum deviation between each sample's cumulative distribution of counts per bin across the entire genome, revealing the greatest extent of distortion or bias relative to the reference distribution. By comparing several test distributions (or genomes) to a single reference, the relative amount of deviation or bias could be determined.

For each genome in this study, the distributions of reads per bin were compared between the five unamplified controls with a KS-test to test for significant variation within the control group. One of the unamplified populations was selected at random to serve as a control, and the distribution of counts per bin from the other four unamplified samples were compared to this. In every case no significant difference (P values ≥ 0.68) was discovered between the control and unamplified replicate populations (please see Additional file 1, Tables S1 and S2). For subsequent comparisons unamplified replicate #3 was chosen (at random) as the control against which the amplified distributions were compared. The D-statistics associated with the control and amplified genomes were then ranked from lowest to highest creating a hierarchy of amplification bias relative to the unamplified reference population.

To ensure that the trends were consistent, the bin size for each genome was increased, effectively decreasing the number of bins and increasing the number of reads per bin, and the process repeated. Four different bin sizes, ranging from 10 bases to $1/100^{th}$ of the full genome size were used for each organism; the results for *Halobacterium* species NRC-1 and *Campylobacter jejuni* are summarized in Tables 6 &7. A graphical summary of the number of reads per 100 base bin for the different sample populations relative to the unamplified control genome can be seen in Figure 5A for *Halobacterium* and Figure 6A for *C. jejuni*.

Discussion

Amplification yield

The yield of the various whole genome amplification techniques, averaged across both genomes, shows clear differences in the amount of DNA generated by the different methodologies. Although other methods exist (PicoGreen, etc), we chose UV spectrophotometry quantification due to the need to detect both single and double stranded DNA, and the fact that all of the DNA had been

Table 5: Raw differences in counts per bin between control and test populations. Numbers in parenthesis are the ratio of the WGA amplified samples normalized to the average deviation in the unamplified replicate populations.

Halobacterium – 100 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	10	10	14	11	43 (3.8)	69 (6.1)	299 (26.6)	1633 (145.2)
Maximum Fold Underamplification	11	12	10	11	17 (1.5)	17 (1.5)	24 (2.2)	74 (6.7)
Maximum Difference in Counts per bin	27	26	31	35	67	101	299	1633
Campylobacter – 100 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	13	13	13	14	16 (1.2)	23 (1.7)	31 (2.3)	2668.5 (201.4)
Maximum Fold Underamplification	11	15	12	14	15 (1.2)	16 (1.2)	19 (1.5)	54 (4.2)
Maximum Difference in Counts per bin	18	17	15	17	23	23	6251	133
Halobacterium – 10 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	11	11	11	11	26 (2.4)	23 (2.1)	52 (4.7)	458 (41.6)
Maximum Fold Underamplification	10	9	9	12	15 (1.5)	15 (1.5)	15 (1.5)	15 (1.5)
Maximum Difference in Counts per bin	10	12	12	12	25	22	51	457
Campylobacter – 10 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	7	9	7	8	9 (1.2)	14 (1.8)	39 (5.0)	991 (127.9)
Maximum Fold Underamplification	8	7	8	7	11 (1.5)	11 (1.5)	11 (1.5)	11 (1.5)
Maximum Difference in Counts per bin	8	8	7	7	10	13	38	990

purified prior to quantitation, thereby removing the excess nucleotides and primers. It is theoretically possible that some of the hexamer primers may have carried through the sodium acetate precipitation step used in the MDA reactions. However, we do not feel that this unduly influenced the yield reported for the MDA reactions as the OD_{260} method was suggested in the GenomiPhi product manual, and the MDA yields we produced were not excessive.

Each reaction started with 25 ng of input DNA, Repli-G generated a 2100 fold amplification of the input DNA, GenomiPhi 640-fold, iPEP 120-fold and DOP amplified the starting material 92-fold. Sharp contrasts are evident between the yields derived from MDA-based (GenomiPhi and Repli-G) amplifications versus PCR-based (iPEP and DOP) approaches; these can be attributed to the differences between the highly processive strand displacement activities of φ29, the polymerase used in both MDA reactions, and the Taq-like enzymes used in the PCR based reactions. Due to the strand displacing capabilities of φ29, MDA reactions do not require repetitive cycles of denaturation and annealing temperatures. In PCR reactions these cycles limit polymerase longevity and activity; the half-life for various Taq polymerases ranges from 30 to 70 minutes at 95°C, resulting in a 50% or greater decrease in enzyme activity at the end of 40 cycles [26]. By utilizing isothermal reactions the MDA methods are able to preserve enzyme functionality for a full 16 hour reaction, and generate substantially more DNA in the process by a hyperbranched mechanism [27] of DNA amplification. It is interesting to note the difference in yield between the MDA methods; it is likely that the use of the KOH alkali denaturation prior to amplification used in the Repli-G process is more efficient at opening potential priming sites than the thermal denaturation used in the GenomiPhi protocol. This is in general agreement with findings that the quality of Sanger-sequence data from MDA-amplified templates improved after alkali denaturation [23], and increased PCR yields following NaOH, as opposed to thermal, denaturation of a high GC genome [28].

Genome coverage

The percent of genome coverage varied considerably by genome and amplification methodology (See Table 1). The unamplified samples, both reference and all unamplified replicates, covered similar expanses of the target genomes, obtaining 84.9 to 85.1% coverage of *Halobacterium* species NRC-1 and 99.6 to 99.7% of *Campylobacter jejuni*. As the *Halobacterium* species NRC-1 genome is 1.54 times larger than *Campylobacter jejuni* (2.56 versus 1.64 MB) it is not surprising that 100,000 reads complete rela-

Table 6: Comparison of *Halobacterium* coverage bias. Kolmogorov-Smirnov comparison of the distributions of reads per bin from an unamplified sample of *Halobacterium* species NRC-1 with an additional unamplified replicate library and libraries amplified with GenomiPhi, Repli-G, PEP and DOP. Bin Size refers to the number of bases comprising each individual bin into which the genome was broken for analysis; 100,000 reads were used for each analysis. Ranked bias was derived from ranked D statistics, lowest to highest.

Unamplified Control versus:	Unamplified Replicate	GenomiPhi	RepliG	PEP	DOP
Bin Size (bp)			10		
Number of Bins			257102		
Number of Reads			100000		
D Statistic	0.001	0.081	0.099	0.108	0.213
P value	1.000	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			100		
Number of Bins			25711		
Number of Reads			100000		
D Statistic	0.003	0.357	0.492	0.495	0.756
P value	0.980	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			2571		
Number of Bins			1001		
Number of Reads			100000		
D Statistic	0.028	0.648	0.717	0.736	0.837
P value	0.791	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			25711		
Number of Bins			100		
Number of Reads			100000		
D Statistic	0.095	0.630	0.650	0.660	0.720
P value	0.737	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1

tively less of the larger *Halobacterium* sequence. However, the substantial difference between the 85% completion of *Halobacterium* and the 99% *Campylobacter jejuni* completion obtained from the unamplified samples requires additional explanation. Analysis of the number of reads per *Halobacterium* chromosome reveals that the number of reads per 100 base bin is disproportionately high for both pNRC100 (12 reads/bin on average) and pNRC200 (8.8 reads per bin) than for the main chromosome NRC1 (an average of 2.2 reads per bin) (See Table 4.). The over-representation of the minichromosome reads even in the unamplified samples may simply reflect a greater abundance of the smaller chromosomes relative to the main chromosome; pNRC100 is a multicopy chromosome [24], and the same might be true for pNRC200. Additionally, a significantly higher percentage of the *Halobacterium* genome is comprised of repeat regions (13.37% versus 2.71% repeat for *C. jejuni*), and the smaller *Halobacterium* chromosomes contain a higher incidence of repeats than the main chromosome (1.4%, 79.4% and 45% repeat for NRC1, pNRC100 and pNRC200 respectively). The repeat regions are disproportionately heavily sequenced relative to their frequency, encompassing 14%, 89% and 75% of the control reads from NRC1, pNRC100 and pNRC200 respectively (Table 4., Figures 3A and 3B). The relative over-coverage of the repeat regions results in fewer of the 100,000 reads covering unique regions, and thus reduces the total coverage of the *Halobacterium* species NRC-1 genome. *Campylobacter jejuni* also experiences relative oversampling of the repeat regions, but with far fewer repeat regions and a smaller total genome than *Halobacterium*, total genome coverage is less affected.

Total coverage of *Halobacterium* dropped significantly when amplified DNA was sequenced, with coverage ranging from 50.1 to 9.7% depending on the amplification process. *C. jejuni* coverage was more complete for most of the amplified material; GenomiPhi, Repli-G and PEP-PCR amplified samples covered 98.9, 99.0 and 94.4% of the genome respectively, while DOP-PCR only covered 17%. The relatively poor coverage of *Halobacterium* by the WGA methods may be due to the presence of the two minichromosomes, and the frequency of repeat regions within them. As mentioned previously, the minichromosomes are more repeat-rich than the main chromosome, there may be relatively more copies of the minichromosomes

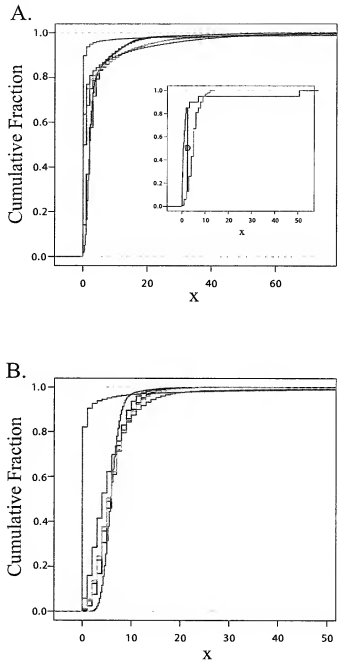

Figure 4

A & B: Empirical cumulative distribution functions. A. Empirical cumulative distribution function (ECDF) depicting the distributions of counts per bin for various control and amplified populations of *Halobacterium* species NRC-1 DNA. The ECDF represents the cumulative distribution of the number of counts per bin, reporting the cumulative proportion of bins with counts equal or less than the value on the X axis. The control cumulative fraction (black) is plotted against all WGA approaches; GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **Inset**. Generic ECDF of two different distributions (red and blue), the D statistic is shown as black vertical line labelled "D". **B** as Figure 4A, but derived from *Campylobacter jejuni*.

Table 7: Comparison of *Campylobacter* coverage bias. Kolmogorov-Smirnov comparison of the distributions of reads per bin from an unamplified sample of *Campylobacter jejuni* with an additional unamplified replicate library and libraries amplified with GenomiPhi, Repli-G, PEP and DOP. Bin Size refers to the number of bases comprising each individual bin into which the genome was broken for analysis; 100,000 reads were used for each analysis. Ranked bias was derived from ranked D statistics, lowest to highest.

Unamplified Control versus:	Unamplified Replicate	GenomiPhi	RepliG	PEP	DOP
Bin Size (bp)			10		
Number of Bins			164149		
Number of Reads			100000		
D Statistic	0.001	0.034	0.029	0.122	0.392
P value	1.000	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	3	4	2	1
Bin Size (bp)			100		
Number of Bins			16415		
Number of Reads			100000		
D Statistic	0.004	0.060	0.079	0.247	0.882
P value	0.997	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			1641		
Number of Bins			1001		
Number of Reads			100000		
D Statistic	0.028	0.214	0.275	0.483	0.915
P value	0.814	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			16413		
Number of Bins			100		
Number of Reads			100000		
D Statistic	0.074	0.248	0.307	0.505	0.842
P value	0.922	0.004	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1

present in the population relative to the main chromosome, and the smaller chromosomes are more likely than the larger main chromosome to have survived the DNA purification process and shipment intact. Coupled with the high incidence of repeat sequences in the minichromosomes, this could cause a disproportionate amplification of the two minichromosomes, and corresponding lower sequence coverage of the remaining chromosome.

Additionally, as the two minichromosomes have lower GC contents than the main chromosome (57.9 and 59.2% for pNRC100 and pNRC200 respectively, relative to 67% for the main chromosome); the possibility of preferential, GC-influenced amplification was investigated. For either genome the mean sequence GC content was not significantly different ($p > 0.05$) between the unamplified control and the replicate populations. The control values for *C. jejuni* were in close agreement with the listed GC content (31% versus 31.15%), while the values obtained for *Halobacterium species NRC-1* were slightly lower than the listed values (61.2% versus 66%), possibly reflecting both the lower proportional coverage of the genome, as well as inherent difficulty in sequencing strands with high degrees of secondary structure.

The mean sequence GC contents for all WGA methodologies were significantly different ($p \leq 0.001$) from the control GC content for both genomes. Deviation from the control GC content was less pronounced for *Campylobacter jejuni*, the low GC genome. Sequences generated by GenomiPhi and Repli-G were roughly 0.2% and 0.5% more GC rich than the control, respectively, while PEP-PCR produced reads that were roughly 2.5% more GC rich than the unamplified control. In contrast, DOP-PCR reads were biased in the opposite direction, 0.3 to 0.4% more AT rich than the control sequence.

The sequences produced by each of the WGA methods were significantly lower in GC than the unamplified *Halobacterium species NRC-1* reference sequence, and the margin of difference was more pronounced than described for *C. jejuni*. GenomiPhi sequences were the closest to the control GC content, with roughly 2.9% lower GC content in the amplified sequence. Repli-G was the next highest,

with roughly 3.8% less GC than the control. The two PCR-based methods generated the most biased sequence in terms of GC content: the mean GC content of PEP-PCR reads were between 4.9 and 5.0% lower than the control, while DOP-PCR reads contained 5.9 to 6.08% less GC on average.

The relative abundance and size of homopolymer stretches in the target genome could possibly influence genome coverage through polymerase slippage in whole genome amplified samples. The frequency of A, C, T and G homopolymers were calculated separately for each genome and sample. The data were $\log_2$ transformed to allow an informative scale and resolution and shown in Figures 1 and 2 for *Halobacterium* and *C. jejuni* respectively. While a thorough analysis of the influence of template composition on sequence coverage and amplification is beyond the scope of this paper, the figures illustrate that homopolymer coverage is similar for most samples across all four nucleotides and both genomes. One exception is found in the DOP-PCR coverage of C and G homopolymers in *C. jejuni* (purple data points in Figure 2B and 2C), where the frequency of the G and C homopolymers is substantially greater than those in the controls and other amplified samples for homopolymers from 6 to 12 nucleotides long.

Sequence coverage bias related to MDA has been reported in the past. Dean *et al.* [29] detected preferential amplification of pUC19 from samples containing both plasmid and bacterial chromosomal DNA. In their case, however, the pUC19 plasmid was only 2.7 kb in circumference, and thus easily circumnavigated by a single φ29 molecule, which generates an average product of 10 kb or greater [15]. In our study, however, the size difference between the main and mini *Halobacterium species NRC-1* chromosomes is less dramatic, and all chromosomes exceed 100 kb in length, rendering it unlikely that polymerase processivity is the root cause of the increased minichromosomes coverage. It is possible, as suggested by Dean *et al.* [29] that the smaller chromosomes are less likely to suffer nicking and subsequent RCA termination than larger chromosomes. As mentioned previously, Paez *et al.* [23] reported selective under-representation and loss from amplified material, Raghunathan *et al.* [9]described bias from 0.1% to 1211% in a qPCR analysis of MDA from single *E. coli* cells, sequence loss related to regional proximity to the ends of both human and yeast chromosomes has also been detected [16], and MDA-induced loci over and under representation as high as 6 fold has been discovered by qPCR [20]. MDA amplification of multiple bacterial species within a single sample revealed preferential amplification of 3 of the 8 species, although all species were represented to some extent [30]. In this study, differences in template size, stochastic effects and other factors were thought to be influential [30].

It is interesting to note that in our study none of the previously mentioned possibilities (repeat regions, chromosome size and GC content) can fully explain the over-amplification of the mini-chromosomes. Should these factors directly influence the relative amplification rate, one would expect pNRC100 to display a higher amplification rate than pNRC200 as pNRC100 is smaller, has a lower GC content, and is substantially more repeat rich (Table 4.). This is not the case; however, as all WGA samples provide higher coverage for pNRC200 than they do for either pNRC1 or pNRC100. As a result, it is impossible to precisely attribute the cause of the bias as uniquely size, copy number, repeat, homopolymer or GC derived, and overamplification appears to be influenced by factors other than those we have discussed here.

Internal controls and run to run reproducibility

Unamplified *Halobacterium* and *C. jejuni* DNA were made into libraries and sequenced as controls both for comparison to the WGA samples and to ensure that any potential systemic bias introduced by the sequencing platform was reproducible from sample to sample. Comparisons between all five unamplified populations from both genomes displayed similar and low count deviation between identical bins (Table 5.) and returned non-significant P-values (P values ≥ 0.68) at all bin sizes (Tables S1 and S2), with similar non-significant P-values between the unamplified reference sample and the unamplified control (See Tables 6 and 7). Similarly, reads per bin from the control plotted against the reference displayed similar location and magnitude of over-represented reads (Figures 5A and 6A, red/black plots), with the previously mentioned over-representation of repeat regions and the smaller minichromosomes (Figures 3A and 3B). Log-log plots of the reads per bin (Figures 5B and 6B, red plots) reveal a strongly linear relationship that indicates that multiple sequencing runs do not induce significant bias in samples generated from the same original genomic library. Having established the unamplified DNA libraries as a baseline, potential bias in the whole genome amplified libraries was quantified.

Bias ranking whole genome amplified template DNA

Regardless of bin size and genome, all whole genome amplified samples displayed significant bias relative to the unamplified reference (See Tables 6 and 7). With one exception the relative amount of bias induced by the various methods remained consistent across both genomes and all bin size resolutions, with GenomiPhi generating the lowest D-statistics, followed by Repli-G, PEP and DOP. The exception was found in the 10 base resolution analysis for *C. jejuni*, where GenomiPhi and Repli-G

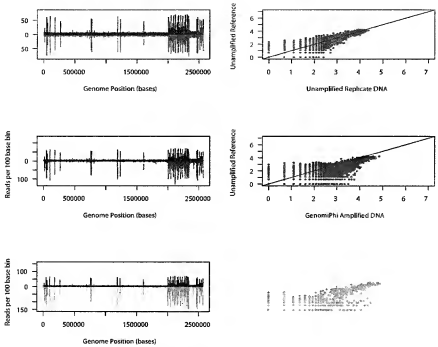

Figure 5

A & B: Comparison of *Halobacterium* sequence coverage. Distribution of sequence coverage from unamplified reference, unamplified replicate and whole genome amplified *Halobacterium* species NRC-1 populations as determined by sequence based karyotyping at 100 base resolution. **A**. Distribution of the counts per bin (y-axis) across the length of the genome (X axis) of test populations relative to the unamplified reference population. The counts per bin from the unamplified reference population are depicted in black above the X axis, counts from the various test populations are inverted and shown below the X axis. The data from the unamplified replicate population is shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B**. Log-log plot of the counts per bin from the unamplified reference population versus the various test populations. The comparison between the reference and the unamplified replicate population is shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. A 45 degree black line is shown for comparison.

BMC Genomics

BioMed Central

Research article

Open Access

Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing

Robert Pinard[1], Alex de Winter[1], Gary J Sarkis[1], Mark B Gerstein[2], Karrie R Tartaro[1], Ramona N Plant[1], Michael Egholm[1], Jonathan M Rothberg[1] and John H Leamon*[1]

Address: [1]454 Life Sciences, 20 Commercial Street, Branford CT 06405, USA and [2]MB&B Department, Yale University, 266 Whitney Ave., New Haven CT 06520, USA

Email: Robert Pinard - robert.pinard@historx.com; Alex de Winter - adewinter@pacificbiosciences.com; Gary J Sarkis - gary.sarkis@ikonisys.com; Mark B Gerstein - mark.gerstein@yale.edu; Karrie R Tartaro - ktartaro@454.com; Ramona N Plant - ramona.2.plant@gsk.com; Michael Egholm - megholm@454.com; Jonathan M Rothberg - jrothberg@454.com; John H Leamon* - jleamon@454.com

* Corresponding author

Published: 23 August 2006

BMC Genomics 2006, 7:216 doi:10.1186/1471-2164-7-216

Received: 24 May 2006

Accepted: 23 August 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/216

© 2006 Pinard et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Whole genome amplification is an increasingly common technique through which minute amounts of DNA can be multiplied to generate quantities suitable for genetic testing and analysis. Questions of amplification-induced error and template bias generated by these methods have previously been addressed through either small scale (SNPs) or large scale (CGH array, FISH) methodologies. Here we utilized whole genome sequencing to assess amplification-induced bias in both coding and non-coding regions of two bacterial genomes. *Halobacterium* species NRC-1 DNA and *Campylobacter jejuni* were amplified by several common, commercially available protocols: multiple displacement amplification, primer extension pre-amplification and degenerate oligonucleotide primed PCR. The amplification-induced bias of each method was assessed by sequencing both genomes in their entirety using the 454 Sequencing System technology and comparing the results with those obtained from unamplified controls.

Results: All amplification methodologies induced statistically significant bias relative to the unamplified control. For the *Halobacterium* species NRC-1 genome, assessed at 100 base resolution, the D-statistics from GenomiPhi-amplified material were 119 times greater than those from unamplified material, 164.0 times greater for Repli-G, 165.0 times greater for PEP-PCR and 252.0 times greater than the unamplified controls for DOP-PCR. For *Campylobacter jejuni*, also analyzed at 100 base resolution, the D-statistics from GenomiPhi-amplified material were 15 times greater than those from unamplified material, 19.8 times greater for Repli-G, 61.8 times greater for PEP-PCR and 220.5 times greater than the unamplified controls for DOP-PCR.

Conclusion: Of the amplification methodologies examined in this paper, the multiple displacement amplification products generated the least bias, and produced significantly higher yields of amplified DNA.

Background

Continued improvement in sequencing quality, combined with increasingly sophisticated bioinformatic analysis of sequence data, has increased the relevance of whole genome sequencing to many fields of biological science including the pharmaceutical industry, agriculture, national defence and medicine [1]. Similarly, the increased availability of sequence data has served to support increasingly complex and informative comparative genomic studies. In both cases, the enhanced relevance and power of genomic comparisons have, in turn, furthered demand for still more sequence data, with the goal of comparing entire genomes. Although high-throughput sequencing methodologies have been developed to accommodate the demand for sequence output, they consume large amounts of a valuable input: genomic DNA. For example, a Taq-man based whole genome association study of 300,000 SNPs would require approximately 9 mg of genomic DNA, more than obtained in routine clinical blood samples [2]. Less input DNA is required for a genome-wide, microarray-based survey restricted to known mutations, but even this would require some form of amplification step [3].

Other applications also place a high demand on potentially scarce DNA. Emerging relationships between specific genotypes and risk factors or disease states have focused attention on DNA samples that are of great medical/scientific importance, but of limited supply, such as tumor samples, lavages, buccal swabs, or samples generated by laser capture microscopy [4]. While laser capture offers single cell accuracy and both lavages and swabs permit minimal patient invasion and discomfort, these methodologies produce far less genomic DNA than less precise, more invasive techniques [5-7]. Some inherently rare samples, such as difficult to culture micro-organisms [8] or genes from an individual bacterium [9] are of great scientific interest, but cannot be sequenced by current technologies without pre-amplification [10,11]. Considerable interest also exists in sequencing low abundance DNA from museum or fossil specimens, although amplification of these samples must address issues of degradation [12] and contamination [13,14]. High rates of consumption, combined with high demand from the scientific community, may result in hard decisions restricting access to these limited or irreplaceable samples.

Whole genome amplification (WGA) can potentially eliminate DNA as a limiting factor for genetic assays. However, in order to fulfil this role, WGA must satisfy some basic requirements. First, the amplification process should be highly accurate, so as to avoid introducing an undue number of errors. Second, amplification should not induce a bias in the distribution of the product DNA. Third, a high amplification factor is required, so that the WGA generates a useful amount of DNA from small starting samples. Finally, the WGA method should be applicable to a wide array of genomes. For maximal efficiency, the WGA protocol would be universally applicable, without need for separate optimization for each sample. In this paper we will address the latter three points – bias, yield, and applicability to two different genomes – leaving the more complicated studies of both amplification fidelity and the sequence-specific causes of bias for another time.

Three primary forms of WGA have been developed: multiple displacement amplification (MDA) [15,16], primer extension preamplification (PEP) [17], and degenerate oligonucleotide primed PCR (DOP) [18]. These WGA methods have been compared in previous papers, but these comparisons have been limited in scale. The authors either scanned individual nucleotide mutations for SNP analysis, or used comparative genomic hybridization (CGH) or fluorescence in situ hybridization (FISH) to scan large regions of the genome [15,19-22]. In the SNP analyses, the comparison is at a high resolution, but small in scope, while for CGH and FISH, the comparison is at a low resolution (since these methods are extremely forgiving of point mutation errors), but large in scope. The different methodology resolutions can therefore report differing levels of bias, as described in a recent study of φ29 fidelity using both direct sequencing and array hybridization of 10000 SNPs [23]. While array-hybridization results revealed whole genome amplification-related loss of 6 regions (approximately 5.56 Mb) and under-representation of another 8 regions, SNP calls from amplified DNA were not statistically different from those of unamplified material [23]. Ideally, any comparison of WGA methods would investigate amplification bias across entire genomes at the highest resolution possible.

For this paper, we have used MDA, PEP and DOP to amplify two bacterial genomes, *Halobacterium* species NRC-1 with a relatively high 66% GC content (derived from the 68% GC main chromosome and two, lower GC content associated minichromosomes, of which at least one, pNRC100, is multicopy [24] – see next section for more detail) and *Campylobacter jejuni* with a single, 31% GC content chromosome. After making libraries of the resultant amplified genomes, we sequenced the libraries using the 454 Sequencing System [25], and enumerated reads initiating within various sized windows of the respective genome, with a maximum resolution of 10 bases across the entire length of the genome. We then conducted sequence-based karyotyping on the amplified and control genomes, and were thus able to generate a high-resolution comparison, encompassing both coding and non-coding regions, of the coverage bias induced by WGA methods across complete bacterial genomes.

Results

Amplification yield

DNA samples were assayed by UV absorption at 260 nm to determine their concentration after amplification. The amount of input DNA was held constant at 25 ng for all methods. Averaged across both genomes, GenomiPhi generated 16.1 μg of DNA, a 640-fold amplification, Repli-G amplified input DNA 2100 fold to 53.6 μg, PEP generated 3.0 μg, a 120-fold increase, and DOP amplified the input DNA 92-fold to 2.3 μg.

Data analysis

For each amplification method, sequencing reads that mapped to the target genome at 95% or greater accuracy were pooled from three or more individual sequencing runs. The percentage of sequences that mapped to the genome varied depending on the amplification method utilized: roughly 60% of the unamplified and MDA amplified reads mapped at 95% accuracy or better, while 40% of the PEP and 20% of the DOP samples mapped to their target genomes. The number of pooled, mapped reads for each of the amplified samples exceeded 150,000, while the number of pooled control reads was approximately 1,500,000 for each genome, reflecting the larger number of samples drawn from these pools to assess variability within the controls.

Analysis populations composed of 100,000 unique reads, their start position (in base pairs), read length (in bases) and their orientation (forward or reverse) on the reference genome were randomly sampled from each of the pooled sequences from amplified genomic material. Five separate analysis populations were generated from each of the control sequence pools to determine the degree of variation within the unamplified reads. Following the generation of the analysis populations, the total genome coverage in bases was determined for each population and both genomes, and the percent of total coverage calculated for each in Table 1.

The GC content of the sequenced reads was determined for each genome and amplification methodology. The FASTA files generated for each of the one hundred thousand reads resulting from each amplification method were analyzed for GC content, and the mean GC content and standard deviation for each method was calculated. Welsh's two sample t-test was employed to compare the mean GC content from each test population against the reference population for each genome. The 95% confi-

Table 1: Comparison of genome coverage. Coverage was derived from the individual sequences generated from either unamplified control or whole genome amplified samples.

Halobacterium species NRC-1			
Sample	**Total Bases Sequenced**	**Nonredundant Bases Sequenced**	**Percent of Genome**
Unamplified Control	9543911	2183656	84.9%
Unamplified Replicate 1	9540329	2188525	85.1%
Unamplified Replicate 2	9540870	2184161	85.0%
Unamplified Replicate 3	9541768	2179647	84.8%
Unamplified Replicate 4	9543517	2183939	84.9%
Averaged Replicates	9541621	2184068	84.9%
GenomiPhi	9736445	1287564	50.1%
Repli-G	9432145	933686	36.3%
PEP-PCR	9410215	921446	35.8%
DOP-PCR	9010086	249571	9.7%
Campylobacter jejuni			
Sample	**Total Bases Sequenced**	**Nonredundant Bases Sequenced**	**Percent of Genome**
Unamplified Control	10605551	1635277	99.6%
Unamplified Replicate 1	10605882	1636507	99.7%
Unamplified Replicate 2	10592192	1636033	99.7%
Unamplified Replicate 3	10603419	1636620	99.7%
Unamplified Replicate 4	10588370	1637090	99.7%
Averaged Replicates	10597466	1636563	99.7%
GenomiPhi	10380921	1623858	98.9%
Repli-G	10939177	1625276	99.0%
PEP-PCR	10313911	1549641	94.4%
DOP-PCR	9605188	278645	17.0%

Table 2: Comparison of *Halobacterium* GC coverage. Mean sequence GC contents, standard deviations, P-values and 95% Confidence limits around differences between the test and unamplified control population means for *Halobacterium* species NRC-1.

Population	P Value	Unamplified Control Mean GC Content	Test Population Mean GC Content	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Unamplified Replicate 1	0.32	61.21 ± 8.6%	61.25 ± 8.6%	-0.11%	0.04%
Unamplified Replicate 2	0.88		61.21 ± 8.6%	-0.07%	0.08%
Unamplified Replicate 3	0.92		61.21 ± 8.6%	-0.07%	0.08%
Unamplified Replicate 4	0.23		61.26 ± 8.6%	-0.12%	0.03%
GenomiPhi	< 0.001		58.29 ± 7.7%	2.85%	2.99%
Repli-G	< 0.001		57.36 ± 7.4%	3.78%	3.92%
PEP-PCR	< 0.001		56.25 ± 7.2%	4.90%	5.04%
DOP-PCR	< 0.001		55.20 ± 7.0%	5.94%	6.08%

dence interval around the difference between the means, and the corresponding P-value were recorded in Tables 2 and 3. The sequencing results for both genomes were also analyzed for the type and size of homopolymers covered in the sequencing and summarized in Figures 1 and 2.

The reference genome was then subdivided into bins of a specific number of bases, and the number of reads that started in each bin was recorded for every population. Additionally, the relationship between coverage depth and the presence of minichromosomes or genomic repeat regions was examined for 100 base genomic bins (See Table 4.). For the purposes of this paper and this sequencing technology, genomic repeats were defined as regions which were 95% identical across 100 bases. The genomic location of the repeat regions relative to the counts per 100 base bins from unamplified Halobacterium and C. jejuni controls is shown in Figure 3A and 3B.

For each sample, the number of reads initiated in each bin was compared to the number found in the same bin of the respective unamplified control. The ratio of the two numbers was computed for each bin, and the maximum (representing over-amplification relative to the control) and minimum ratio (representing under-amplification or sequence loss relative to the control) were recorded in Table 5. As an identical number of reads were used for each sample, the average numbers of reads obtained per bin were identical for each treatment, requiring a more sophisticated statistical assessment to accurately assess potential bias.

The empirical cumulative frequency distribution (ECDF) of the reads per bin for *Halobacterium* and *C. jejuni* are shown in Figures 4A and 4B respectively. The ECDF represents the cumulative distribution of the number of counts per bin, reporting the cumulative proportion of bins with counts equal or less than the value on the X axis. It was expected that the counts per bin would follow a Poisson distribution, and some bins in each sample would doubtlessly contain outliers. To address relative bias we wanted to compare differences between the read distributions obtained for each sample, rather than comparing each sample to a model distribution. As a result, the non-parametric, distribution-free Kolmogorov-Smirnov test (KS-test) with its associated D-statistic was used for subsequent analysis.

The Kolmogorov-Smirnov test (KS-test) and D-statistic

The KS-test was employed to assess the deviation between the number of counts per bin that occurred between two genomes (one test, one reference), using the maximal deviation (D-statistic) between the cumulative count distributions of the populations being compared (for a visual

Table 3: Comparison of *Campylobacter jejuni* GC coverage. Mean sequence GC contents, standard deviations, P-values and 95% Confidence limits around differences between the test and unamplified reference population means for *Campylobacter jejuni*.

Population	P Value	Unamplified Control Mean GC Content	Test Population Mean GC Content	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Unamplified Replicate 1	0.65	31.15 ± 6.7%	31.16 ± 6.7%	-0.07%	0.05%
Unamplified Replicate 2	0.75		31.16 ± 6.7%	-0.07%	0.05%
Unamplified Replicate 3	0.83		31.15 ± 6.7%	-0.07%	0.05%
Unamplified Replicate 4	0.75		31.16 ± 6.7%	-0.07%	0.05%
GenomiPhi	< 0.001		31.36 ± 6.8%	-0.28%	-0.16%
Repli-G	< 0.001		31.60 ± 6.7%	-0.52%	-0.40%
PEP-PCR	< 0.001		33.92 ± 8.1%	-2.84%	-2.71%
DOP-PCR	< 0.001		30.79 ± 5.9%	0.30%	0.41%

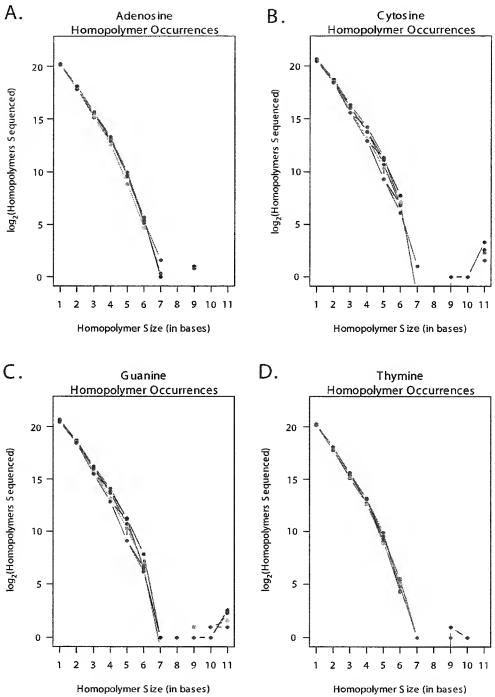

Figure 1

A – D: Hompolymer coverage in *Halobacterium* sequence reads. A. A $\log_2$plot illustrating the total number of adenosine homopolymers sequenced in reads generated by control and amplified populations of *Halobacterium* species NRC-1 DNA. The data from the unamplified replicate population are shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B**. As Figure 1A., but for Cytosine homopolymers. **C**. As Figure 1A., but for Guanine homopolymers. **D**. As Figure 1A., but for Thymine homopolymers.

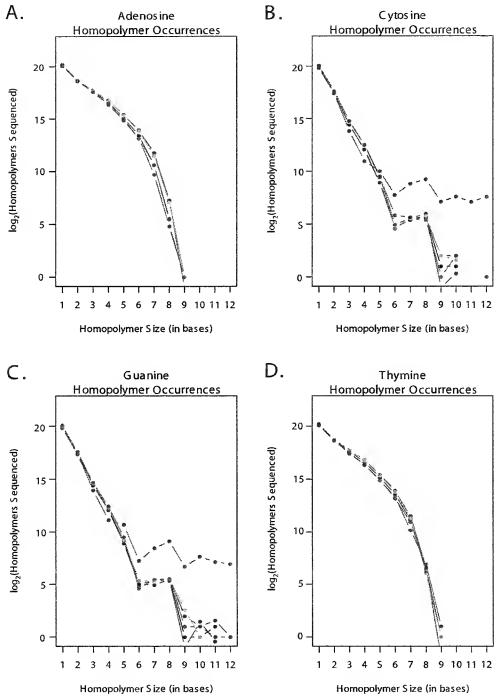

Figure 2

A – D: Hompolymer coverage in *Campylobacter jejuni* sequence reads. A. A $\log_2$ plot illustrating the total number of adenosine homopolymers sequenced in reads generated by control and amplified populations of *Campylobacter jejuni* DNA. The data from the unamplified replicate population are shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B.** As Figure 2A., but for Cytosine homopolymers. **C.** As Figure 2A., but for Guanine homopolymers. **D.** As Figure 2A., but for Thymine homopolymers.

Table 4: Comparison of chromosomal coverage. Sequence coverage per 100 base bin is shown for both repeat and unique regions from unamplified control or whole genome amplified samples.

Halobacterium species NRC-1 main (NRC1) and minichromomes (pNRC100 and pNRC200)

Chromosome	Sample	Reads per unique bin	Reads per repeat bin	Percent chromosomal reads/genome	Percent unique chromosomal reads/genome	Percent repeat chromosomal reads/genome	Percent unique reads/ chromosomal reads	Percent repeat reads/ chromosomal reads
NRC1 2.01 Mb 68% GC 1.4% repeats	Control	1.91	22.38	44%	38%	6%	86%	14%
	Average Unamplified Replicates	1.92	22.18	44%	38%	6%	86%	14%
	GenomiPhi	0.83	30.92	25%	16%	9%	66%	35%
	Repli-G	0.55	36.23	21%	11%	10%	52%	48%
	PEP-PCR	0.69	28.67	22%	14%	8%	63%	37%
	DOP-PCR	0.38	28.07	15%	8%	8%	49%	51%
pNRC100 191Kb 57.9% GC 79.4% repeats	Control	6.42	13.84	24%	3%	21%	11%	89%
	Average Unamplified Replicates	6.51	13.77	23%	3%	21%	11%	89%
	GenomiPhi	9.47	18.64	32%	4%	28%	12%	88%
	Repli-G	9.81	19.68	34%	4%	30%	12%	88%
	PEP-PCR	11.09	20.51	35%	4%	31%	13%	88%
	DOP-PCR	2.59	20.19	32%	1%	31%	3%	97%
pNRC200 365 Kb 59.2% GC 45.0% repeats	Control	3.78	14.97	32%	8%	25%	24%	76%
	Average Unamplified Replicates	3.85	14.89	32%	8%	24%	24%	76%
	GenomiPhi	4.86	20.19	43%	10%	33%	23%	77%
	Repli-G	4.54	21.92	45%	9%	36%	20%	80%
	PEP-PCR	3.76	21.40	43%	8%	35%	18%	82%
	DOP-PCR	7.65	22.92	53%	15%	38%	29%	71%

Campylobacter jejuni single chromosome

Chromosome	Sample	Reads per unique bin	Reads per repeat bin	Percent chromosomal reads/genome	Percent unique chromosomal reads/genome	Percent repeat chromosomal reads/genome	Percent unique reads/ chromosomal reads	Percent repeat reads/ chromosomal reads
C. Jejuni 1.64 Mb 31% GC 2.71% repeat	Control	5.90	12.94	N/A	94%	6%	N/A	N/A
	Average Unamplified Replicates	5.90	12.97		94%	6%		
	GenomiPhi	5.96	10.81		95%	5%		
	Repli-G	5.87	13.85		94%	6%		
	PEP-PCR	5.35	32.42		86%	15%		
	DOP-PCR	6.12	4.91		98%	2%		

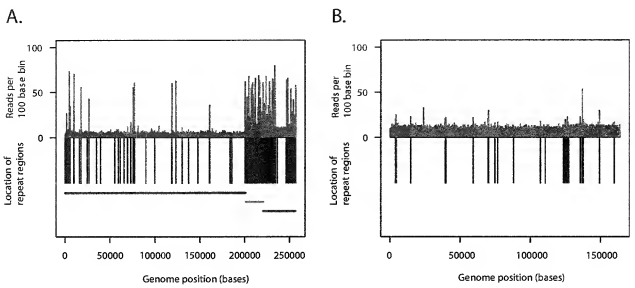

Figure 3

A & B: Comparison of control sequence coverage versus repeat region location. Distribution of sequence coverage from unamplified control *Halobacterium* species NRC-1 population as determined by sequence based karyotyping at 100 base resolution relative to repeat region and chromosome location. **A.** Counts per bin are displayed above the X axis in red, repeat regions are shown below the X axis in black. The X axis displays the length of the genome in bases. The relative location of the three *Halobacterium* chromosomes are shown by the horizontal bars below the X axis, NRC1 is green, pNRC100 is gold, and pNRC200 is purple. **B.** As for Figure 3A, but for *Campylobacter jejuni*. No chromosome bars are included as *C. jejuni* is comprised of a single chromosome.

demonstration of the D-statistic, please see the inset on Figure 4A). The D-statistic for each comparison reported the maximum deviation between each sample's cumulative distribution of counts per bin across the entire genome, revealing the greatest extent of distortion or bias relative to the reference distribution. By comparing several test distributions (or genomes) to a single reference, the relative amount of deviation or bias could be determined.

For each genome in this study, the distributions of reads per bin were compared between the five unamplified controls with a KS-test to test for significant variation within the control group. One of the unamplified populations was selected at random to serve as a control, and the distribution of counts per bin from the other four unamplified samples were compared to this. In every case no significant difference (P values ≥ 0.68) was discovered between the control and unamplified replicate populations (please see Additional file 1, Tables S1 and S2). For subsequent comparisons unamplified replicate #3 was chosen (at random) as the control against which the amplified distributions were compared. The D-statistics associated with the control and amplified genomes were then ranked from lowest to highest creating a hierarchy of amplification bias relative to the unamplified reference population.

To ensure that the trends were consistent, the bin size for each genome was increased, effectively decreasing the number of bins and increasing the number of reads per bin, and the process repeated. Four different bin sizes, ranging from 10 bases to 1/100th of the full genome size were used for each organism; the results for *Halobacterium* species NRC-1 and *Campylobacter jejuni* are summarized in Tables 6 &7. A graphical summary of the number of reads per 100 base bin for the different sample populations relative to the unamplified control genome can be seen in Figure 5A for *Halobacterium* and Figure 6A for *C. jejuni*.

Discussion

Amplification yield

The yield of the various whole genome amplification techniques, averaged across both genomes, shows clear differences in the amount of DNA generated by the different methodologies. Although other methods exist (PicoGreen, etc), we chose UV spectrophotometry quantification due to the need to detect both single and double stranded DNA, and the fact that all of the DNA had been

Table 5: Raw differences in counts per bin between control and test populations. Numbers in parenthesis are the ratio of the WGA amplified samples normalized to the average deviation in the unamplified replicate populations.

Halobacterium – 100 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	10	10	14	11	43 (3.8)	69 (6.1)	299 (26.6)	1633 (145.2)
Maximum Fold Underamplification	11	12	10	11	17 (1.5)	17 (1.5)	24 (2.2)	74 (6.7)
Maximum Difference in Counts per bin	27	26	31	35	67	101	299	1633
Campylobacter – 100 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	13	13	13	14	16 (1.2)	23 (1.7)	31 (2.3)	2668.5 (201.4)
Maximum Fold Underamplification	11	15	12	14	15 (1.2)	16 (1.2)	19 (1.5)	54 (4.2)
Maximum Difference in Counts per bin	18	17	15	17	23	23	6251	133
Halobacterium – 10 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	11	11	11	11	26 (2.4)	23 (2.1)	52 (4.7)	458 (41.6)
Maximum Fold Underamplification	10	9	9	12	15 (1.5)	15 (1.5)	15 (1.5)	15 (1.5)
Maximum Difference in Counts per bin	10	12	12	12	25	22	51	457
Campylobacter – 10 base bins								
Versus Unamplified Control	Replicate 1	Replicate 2	Replicate 3	Replicate 4	GenomiPhi	Repli-G	PEP-PCR	DOP-PCR
Maximum Fold Overamplification	7	9	7	8	9 (1.2)	14 (1.8)	39 (5.0)	991 (127.9)
Maximum Fold Underamplification	8	7	8	7	11 (1.5)	11 (1.5)	11 (1.5)	11 (1.5)
Maximum Difference in Counts per bin	8	8	7	7	10	13	38	990

purified prior to quantitation, thereby removing the excess nucleotides and primers. It is theoretically possible that some of the hexamer primers may have carried through the sodium acetate precipitation step used in the MDA reactions. However, we do not feel that this unduly influenced the yield reported for the MDA reactions as the OD_{260} method was suggested in the GenomiPhi product manual, and the MDA yields we produced were not excessive.

Each reaction started with 25 ng of input DNA, Repli-G generated a 2100 fold amplification of the input DNA, GenomiPhi 640-fold, iPEP 120-fold and DOP amplified the starting material 92-fold. Sharp contrasts are evident between the yields derived from MDA-based (GenomiPhi and Repli-G) amplifications versus PCR-based (iPEP and DOP) approaches; these can be attributed to the differences between the highly processive strand displacement activities of φ29, the polymerase used in both MDA reactions, and the Taq-like enzymes used in the PCR based reactions. Due to the strand displacing capabilities of φ29, MDA reactions do not require repetitive cycles of denaturation and annealing temperatures. In PCR reactions these cycles limit polymerase longevity and activity; the half-life for various Taq polymerases ranges from 30 to 70 minutes at 95°C, resulting in a 50% or greater decrease in enzyme activity at the end of 40 cycles [26]. By utilizing isothermal reactions the MDA methods are able to preserve enzyme functionality for a full 16 hour reaction, and generate substantially more DNA in the process by a hyperbranched mechanism [27] of DNA amplification. It is interesting to note the difference in yield between the MDA methods; it is likely that the use of the KOH alkali denaturation prior to amplification used in the Repli-G process is more efficient at opening potential priming sites than the thermal denaturation used in the GenomiPhi protocol. This is in general agreement with findings that the quality of Sanger-sequence data from MDA-amplified templates improved after alkali denaturation [23], and increased PCR yields following NaOH, as opposed to thermal, denaturation of a high GC genome [28].

Genome coverage

The percent of genome coverage varied considerably by genome and amplification methodology (See Table 1). The unamplified samples, both reference and all unamplified replicates, covered similar expanses of the target genomes, obtaining 84.9 to 85.1% coverage of *Halobacterium* species NRC-1 and 99.6 to 99.7% of *Campylobacter jejuni*. As the *Halobacterium* species NRC-1 genome is 1.54 times larger than *Campylobacter jejuni* (2.56 versus 1.64 MB) it is not surprising that 100,000 reads complete rela-

Table 6: Comparison of *Halobacterium* coverage bias. Kolmogorov-Smirnov comparison of the distributions of reads per bin from an unamplified sample of *Halobacterium* species NRC-1 with an additional unamplified replicate library and libraries amplified with GenomiPhi, Repli-G, PEP and DOP. Bin Size refers to the number of bases comprising each individual bin into which the genome was broken for analysis; 100,000 reads were used for each analysis. Ranked bias was derived from ranked D statistics, lowest to highest.

Unamplified Control versus:	Unamplified Replicate	GenomiPhi	RepliG	PEP	DOP
Bin Size (bp)			10		
Number of Bins			257102		
Number of Reads			100000		
D Statistic	0.001	0.081	0.099	0.108	0.213
P value	1.000	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			100		
Number of Bins			25711		
Number of Reads			100000		
D Statistic	0.003	0.357	0.492	0.495	0.756
P value	0.980	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			2571		
Number of Bins			1001		
Number of Reads			100000		
D Statistic	0.028	0.648	0.717	0.736	0.837
P value	0.791	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			25711		
Number of Bins			100		
Number of Reads			100000		
D Statistic	0.095	0.630	0.650	0.660	0.720
P value	0.737	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1

tively less of the larger *Halobacterium* sequence. However, the substantial difference between the 85% completion of *Halobacterium* and the 99% *Campylobacter jejuni* completion obtained from the unamplified samples requires additional explanation. Analysis of the number of reads per *Halobacterium* chromosome reveals that the number of reads per 100 base bin is disproportionately high for both pNRC100 (12 reads/bin on average) and pNRC200 (8.8 reads per bin) than for the main chromosome NRC1 (an average of 2.2 reads per bin) (See Table 4.). The over-representation of the minichromosome reads even in the unamplified samples may simply reflect a greater abundance of the smaller chromosomes relative to the main chromosome; pNRC100 is a multicopy chromosome [24], and the same might be true for pNRC200. Additionally, a significantly higher percentage of the *Halobacterium* genome is comprised of repeat regions (13.37% versus 2.71% repeat for *C. jejuni*), and the smaller *Halobacterium* chromosomes contain a higher incidence of repeats than the main chromosome (1.4%, 79.4% and 45% repeat for NRC1, pNRC100 and pNRC200 respectively). The repeat regions are disproportionately heavily sequenced relative to their frequency, encompassing 14%, 89% and 75% of the control reads from NRC1, pNRC100 and pNRC200 respectively (Table 4., Figures 3A and 3B). The relative over-coverage of the repeat regions results in fewer of the 100,000 reads covering unique regions, and thus reduces the total coverage of the *Halobacterium* species NRC-1 genome. *Campylobacter jejuni* also experiences relative oversampling of the repeat regions, but with far fewer repeat regions and a smaller total genome than *Halobacterium*, total genome coverage is less affected.

Total coverage of *Halobacterium* dropped significantly when amplified DNA was sequenced, with coverage ranging from 50.1 to 9.7% depending on the amplification process. *C. jejuni* coverage was more complete for most of the amplified material; GenomiPhi, Repli-G and PEP-PCR amplified samples covered 98.9, 99.0 and 94.4% of the genome respectively, while DOP-PCR only covered 17%. The relatively poor coverage of *Halobacterium* by the WGA methods may be due to the presence of the two minichromosomes, and the frequency of repeat regions within them. As mentioned previously, the minichromosomes are more repeat-rich than the main chromosome, there may be relatively more copies of the minichromosomes

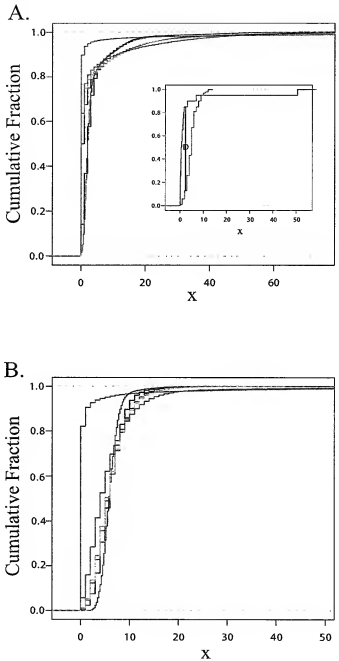

Figure 4

A & B: Empirical cumulative distribution functions. A. Empirical cumulative distribution function (ECDF) depicting the distributions of counts per bin for various control and amplified populations of *Halobacterium* species NRC-1 DNA. The ECDF represents the cumulative distribution of the number of counts per bin, reporting the cumulative proportion of bins with counts equal or less than the value on the X axis. The control cumulative fraction (black) is plotted against all WGA approaches; GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **Inset.** Generic ECDF of two different distributions (red and blue), the D statistic is shown as black vertical line labelled "D". **B** as Figure 4A, but derived from *Campylobacter jejuni*.

Table 7: Comparison of *Campylobacter* coverage bias. Kolmogorov-Smirnov comparison of the distributions of reads per bin from an unamplified sample of *Campylobacter jejuni* with an additional unamplified replicate library and libraries amplified with GenomiPhi, Repli-G, PEP and DOP. Bin Size refers to the number of bases comprising each individual bin into which the genome was broken for analysis; 100,000 reads were used for each analysis. Ranked bias was derived from ranked D statistics, lowest to highest.

Unamplified Control versus:	Unamplified Replicate	GenomiPhi	RepliG	PEP	DOP
Bin Size (bp)			10		
Number of Bins			164149		
Number of Reads			100000		
D Statistic	0.001	0.034	0.029	0.122	0.392
P value	1.000	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	3	4	2	1
Bin Size (bp)			100		
Number of Bins			16415		
Number of Reads			100000		
D Statistic	0.004	0.060	0.079	0.247	0.882
P value	0.997	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			1641		
Number of Bins			1001		
Number of Reads			100000		
D Statistic	0.028	0.214	0.275	0.483	0.915
P value	0.814	0.000	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1
Bin Size (bp)			16413		
Number of Bins			100		
Number of Reads			100000		
D Statistic	0.074	0.248	0.307	0.505	0.842
P value	0.922	0.004	0.000	0.000	0.000
Ranked Bias (5 is lowest)	5	4	3	2	1

present in the population relative to the main chromosome, and the smaller chromosomes are more likely than the larger main chromosome to have survived the DNA purification process and shipment intact. Coupled with the high incidence of repeat sequences in the minichromosomes, this could cause a disproportionate amplification of the two minichromosomes, and corresponding lower sequence coverage of the remaining chromosome.

Additionally, as the two minichromosomes have lower GC contents than the main chromosome (57.9 and 59.2% for pNRC100 and pNRC200 respectively, relative to 67% for the main chromosome); the possibility of preferential, GC-influenced amplification was investigated. For either genome the mean sequence GC content was not significantly different ($p > 0.05$) between the unamplified control and the replicate populations. The control values for *C. jejuni* were in close agreement with the listed GC content (31% versus 31.15%), while the values obtained for *Halobacterium species NRC-1* were slightly lower than the listed values (61.2% versus 66%), possibly reflecting both the lower proportional coverage of the genome, as well as inherent difficulty in sequencing strands with high degrees of secondary structure.

The mean sequence GC contents for all WGA methodologies were significantly different ($p \leq 0.001$) from the control GC content for both genomes. Deviation from the control GC content was less pronounced for *Campylobacter jejuni*, the low GC genome. Sequences generated by GenomiPhi and Repli-G were roughly 0.2% and 0.5% more GC rich than the control, respectively, while PEP-PCR produced reads that were roughly 2.5% more GC rich than the unamplified control. In contrast, DOP-PCR reads were biased in the opposite direction, 0.3 to 0.4% more AT rich than the control sequence.

The sequences produced by each of the WGA methods were significantly lower in GC than the unamplified *Halobacterium species NRC-1* reference sequence, and the margin of difference was more pronounced than described for *C. jejuni*. GenomiPhi sequences were the closest to the control GC content, with roughly 2.9% lower GC content in the amplified sequence. Repli-G was the next highest,

with roughly 3.8% less GC than the control. The two PCR-based methods generated the most biased sequence in terms of GC content: the mean GC content of PEP-PCR reads were between 4.9 and 5.0% lower than the control, while DOP-PCR reads contained 5.9 to 6.08% less GC on average.

The relative abundance and size of homopolymer stretches in the target genome could possibly influence genome coverage through polymerase slippage in whole genome amplified samples. The frequency of A, C, T and G homopolymers were calculated separately for each genome and sample. The data were $\log_2$ transformed to allow an informative scale and resolution and shown in Figures 1 and 2 for *Halobacterium* and *C. jejuni* respectively. While a thorough analysis of the influence of template composition on sequence coverage and amplification is beyond the scope of this paper, the figures illustrate that homopolymer coverage is similar for most samples across all four nucleotides and both genomes. One exception is found in the DOP-PCR coverage of C and G homopolymers in *C. jejuni* (purple data points in Figure 2B and 2C), where the frequency of the G and C homopolymers is substantially greater than those in the controls and other amplified samples for homopolymers from 6 to 12 nucleotides long.

Sequence coverage bias related to MDA has been reported in the past. Dean *et al.* [29] detected preferential amplification of pUC19 from samples containing both plasmid and bacterial chromosomal DNA. In their case, however, the pUC19 plasmid was only 2.7 kb in circumference, and thus easily circumnavigated by a single φ29 molecule, which generates an average product of 10 kb or greater [15]. In our study, however, the size difference between the main and mini *Halobacterium species NRC-1* chromosomes is less dramatic, and all chromosomes exceed 100 kb in length, rendering it unlikely that polymerase processivity is the root cause of the increased minichromosomes coverage. It is possible, as suggested by Dean *et al.* [29] that the smaller chromosomes are less likely to suffer nicking and subsequent RCA termination than larger chromosomes. As mentioned previously, Paez *et al.* [23] reported selective under-representation and loss from amplified material, Raghunathan *et al.* [9]described bias from 0.1% to 1211% in a qPCR analysis of MDA from single *E. coli* cells, sequence loss related to regional proximity to the ends of both human and yeast chromosomes has also been detected [16], and MDA-induced loci over and under representation as high as 6 fold has been discovered by qPCR [20]. MDA amplification of multiple bacterial species within a single sample revealed preferential amplification of 3 of the 8 species, although all species were represented to some extent [30]. In this study, differences in template size, stochastic effects and other factors were thought to be influential [30].

It is interesting to note that in our study none of the previously mentioned possibilities (repeat regions, chromosome size and GC content) can fully explain the over-amplification of the mini-chromosomes. Should these factors directly influence the relative amplification rate, one would expect pNRC100 to display a higher amplification rate than pNRC200 as pNRC100 is smaller, has a lower GC content, and is substantially more repeat rich (Table 4.). This is not the case; however, as all WGA samples provide higher coverage for pNRC200 than they do for either pNRC1 or pNRC100. As a result, it is impossible to precisely attribute the cause of the bias as uniquely size, copy number, repeat, homopolymer or GC derived, and overamplification appears to be influenced by factors other than those we have discussed here.

Internal controls and run to run reproducibility

Unamplified *Halobacterium* and *C. jejuni* DNA were made into libraries and sequenced as controls both for comparison to the WGA samples and to ensure that any potential systemic bias introduced by the sequencing platform was reproducible from sample to sample. Comparisons between all five unamplified populations from both genomes displayed similar and low count deviation between identical bins (Table 5.) and returned non-significant P-values (P values ≥ 0.68) at all bin sizes (Tables S1 and S2), with similar non-significant P-values between the unamplified reference sample and the unamplified control (See Tables 6 and 7). Similarly, reads per bin from the control plotted against the reference displayed similar location and magnitude of over-represented reads (Figures 5A and 6A, red/black plots), with the previously mentioned over-representation of repeat regions and the smaller minichromosomes (Figures 3A and 3B). Log-log plots of the reads per bin (Figures 5B and 6B, red plots) reveal a strongly linear relationship that indicates that multiple sequencing runs do not induce significant bias in samples generated from the same original genomic library. Having established the unamplified DNA libraries as a baseline, potential bias in the whole genome amplified libraries was quantified.

Bias ranking whole genome amplified template DNA

Regardless of bin size and genome, all whole genome amplified samples displayed significant bias relative to the unamplified reference (See Tables 6 and 7). With one exception the relative amount of bias induced by the various methods remained consistent across both genomes and all bin size resolutions, with GenomiPhi generating the lowest D-statistics, followed by Repli-G, PEP and DOP. The exception was found in the 10 base resolution analysis for *C. jejuni*, where GenomiPhi and Repli-G

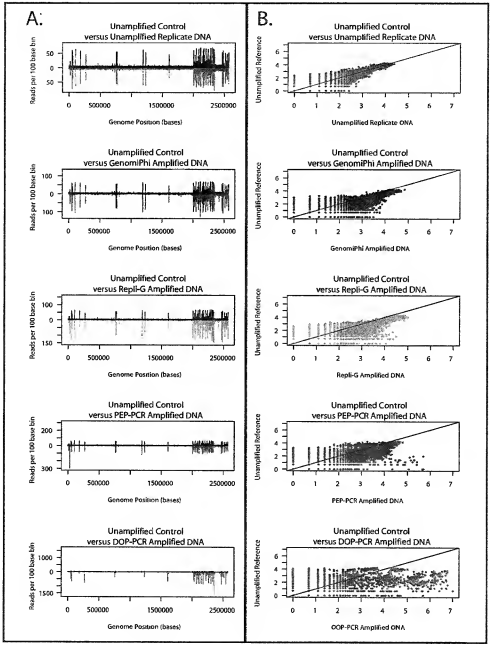

Figure 5

A & B: Comparison of *Halobacterium* sequence coverage. Distribution of sequence coverage from unamplified reference, unamplified replicate and whole genome amplified *Halobacterium* species NRC-1 populations as determined by sequence based karyotyping at 100 base resolution. **A.** Distribution of the counts per bin (y-axis) across the length of the genome (X axis) of test populations relative to the unamplified reference population. The counts per bin from the unamplified reference population are depicted in black above the X axis, counts from the various test populations are inverted and shown below the X axis. The data from the unamplified replicate population is shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B.** Log-log plot of the counts per bin from the unamplified reference population versus the various test populations. The comparison between the reference and the unamplified replicate population is shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. A 45 degree black line is shown for comparison.

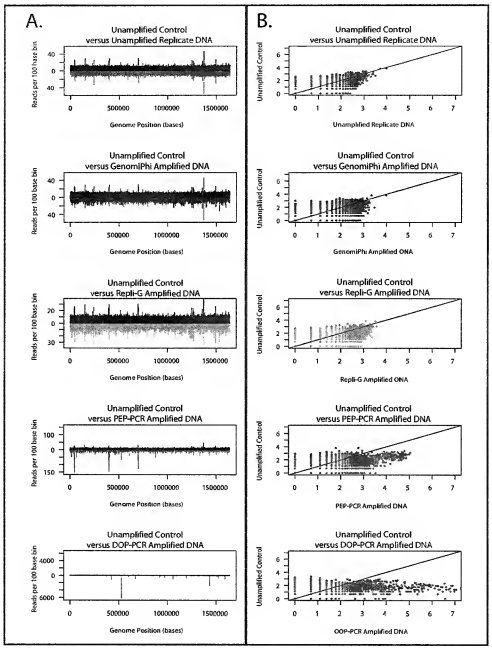

Figure 6

A & B: Comparison of *Campylobacter* sequence coverage. Distribution of sequence coverage from unamplified reference, unamplified replicate and whole genome amplified *Campylobacter jejuni* populations as determined by sequence based karyotyping at 100 base resolution. **A**. Distribution of the counts per bin (y-axis) across the length of the genome (X axis) of test populations relative to the unamplified reference population. The counts per bin from the unamplified reference population are depicted in black above the X axis, counts from the various test populations are inverted and shown below the X axis. The data from the unamplified replicate population is shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. **B**. Log-log plot of the counts per bin from the unamplified reference population versus the various test populations. The comparison between the reference and the unamplified replicate population is shown in red, GenomiPhi in blue, Repli-G in orange, PEP in green and DOP in purple. A 45 degree black line is shown for comparison.

Table 8: WGA-induced bias relative to control populations. KS-test D-statistic ratios relative to the unamplified replicate at increasing bin sizes.

Holobacterium species NRC-1				
Bin Size (bases)	**GenomiPhi**	**RepliG**	**PEP**	**DOP**
10	81.0	99.0	108.0	213.0
100	119.0	164.0	165.0	252.0
2571	23.1	25.6	26.3	29.9
25711	6.6	6.8	7.0	7.6
Compylobacter jejuni				
Bin Size (bases)	**GenomiPhi**	**RepliG**	**PEP**	**DOP**
10	34.0	29.0	122.0	392.0
100	15.0	19.8	61.8	220.5
1641	7.6	9.8	17.2	32.7
16413	3.4	4.2	6.8	11.4

exchanged position; while significant, the difference between the Repli-G and GenomiPhi samples in this instance was slight (0.005). Increased bin size elevated the magnitude of the D-statistics for all test and control populations, although the unamplified control D-statistics increased more rapidly with bin size than did the test populations (Table 8), effectively decreasing the bias as bin size increased. This, however, is merely the result of smoothing a constant number of reads per population across a diminishing number of bins, and does not reflect any amplification-specific size related effect.

The deviation in counts per bin (Table 5) and the D-statistics were 2- to 10-fold lower for *C. jejuni* than for *Halobacterium* for all amplified populations, although the biases between the amplified *C. jejuni* populations were more extreme (Table 9). This may reflect the fact that *C. jejuni*, a smaller circular genome, is better suited to the isothermal, highly processive MDA approaches than the cyclically denatured, less-processive PCR-based methods. Similar to the genome coverage results, the D-statistics indicated that the presence of mini-chromosomes and/or high GC content in *Halobacterium* challenged all amplification techniques and reduced the advantage from the MDA process.

Table 9: Maximum and minimum bias relative to control populations. Minimum and maximum fold differences in KS-test D-statistic values between amplified samples for all bin sizes.

Holobacterium species NRC-1			
	Repli-G	**PEP**	**DOP**
GenomiPhi	1.03 to 1.37	1.04 to 1.38	1.14 to 2.62
Repli-G		1.01 to 1.09	1.11 to 2.15
PEP			1.09 to 1.98
Compylobacter jejuni			
	Repli-G	**PEP**	**DOP**
GenomiPhi	0.85 to 1.30	2.04 to 4.09	3.4 to 14.58
Repli-G		1.64 to 4.28	2.74 to 13.7
PEP			1.67 to 3.57

Multiple displacement amplification methods

The two Multiple Displacement Amplification (MDA) methods, GenomiPhi and Repli-G, introduced the lowest amplification bias, inducing the least distortion in counts per bin across the length of the genome and slight decreases in slope below the 45 degree line in the log-log plot (Figures 5 and 6, blue and orange plots respectively). The maximum deviation from the reference distribution at 10 base resolution was 81 and 34 fold greater than the unamplified control for GenomiPhi on *Halobacterium* and *C. jejuni* respectively, and 99 and 29 fold greater than *Halobacterium* and *C. jejuni* control for Repli-G (Table 8). Differences between the MDA methods were slight, particularly in light of the aforementioned reversal in bias rank in the *C. jejuni* 10 base bin analysis. Regardless of bin size, the D-statistic for Repli-G was 1.03 to 1.37 times larger than that from GenomiPhi for *Halobacterium* and 0.85 to 1.30 times greater for *C. jejuni* (Table 9).

The bias estimates we derived for MDA are comparable to those reported in the literature. The crude metric of counts per bin (Table 5) yields estimates of 2.4 (10 base bin) to 3.8 fold (100 base bin) overamplification relative to the average counts per bin for the unamplified replicates for *Halobacterium* when GenomiPhi was used, and 2.1 to 6.1 fold for Repli-G. Raw count bias in *C. jejuni* was 1.2 fold for both bin sizes for GenomiPhi and 1.7 (100 base bins) to 1.8 (10 base bins) for Repli-G. The biases estimated for each genome by D-statistic are smaller but still comparable to the 0.5- to 3-fold biases reported by Dean *et al.* [15] and Hosono *et al.* [20] as well as the 0.75 to 1.35-fold biases for multi-cell samples in Raghunathan's study [9]. This is particularly interesting as the studies relied upon gene locus sequencing or qPCR quantitation, which focuses exclusively on coding regions of the genome, while our whole genome sequencing included both coding and non-coding regions, yet still reached similar estimates of bias.

PCR-based amplification methods

PCR-based amplification methods were more biased than MDA methods for all bin sizes and both genomes, although the PEP method was less biased than the DOP amplified samples. Visually, the counts per bin are obviously distorted around the repeat regions for PEP ampli-

fied material (Figures 5A &6A, green plot), and even more drastically distorted for DOP samples (Figures 5A &6A, purple plot). Log-log plots display a distinct non-linear relationship between reference and amplified samples, with the data cloud shifting to the right for both amplification methods, with more skew evident in DOP than PEP (Figures 5B &6B). At a 10 base resolution, D-statistics from PEP amplified *Halobacterium* and *C. jejuni* were 108 and 122 times greater than the reference value respectively, while D-statistics from DOP amplified material were respectively 213 and 392 times larger than the *Halobacterium* and *C. jejuni* reference. In comparisons between the PCR-based methods, the maximal deviation from the reference for the PEP amplified *Halobacterium* sample was 1.09 to 1.98 times lower than that amplified by DOP; the D-statistic from the *C. jejuni* sample was 1.67 to 3.57 fold lower with PEP relative to DOP (See Table 9.).

Differing results between the PCR-based amplification methods can be partially attributed to the primer composition in each reaction. Both MDA and PEP amplification used random bases (6 bases for MDA, 15 bases for PEP), while DOP employed a 22 base primer containing a random 6-mer flanked with specific 10-mer 5' and 6-mer 3' sequences. As amplification events were initiated from the annealed specific 3' region of the primer, this resulted in preferential amplification or over-amplification of genomic regions that permitted successful primer binding. Similarly, regions where the primer hybridized poorly were amplified infrequently. Additionally, PCR-driven competition for primers, polymerase, etc. may prevent any significant representation of the infrequently primed amplicons in the final amplified pool. These factors resulted in over-representation of genome regions containing the 3' primer sequence, and was clearly evident in the overamplification spike around base position 540,000 (which does not coincide with a repeat region in the unamplified references) in *C. jejuni* genome. In the course of sequencing the DOP libraries, it was noted that sequences containing the DOP 5' specific decamer (5' CCGACTCGAG 3') also contained the exact 3' hexamer (5' ATGTGG 3') 88% of the time, and matched 5 of the 6 bases at the 3' end 95% of the time. Despite the low stringency of the DOP thermocycling protocol, the DOP primer annealed primarily to regions with at least 5 of 6 bases of homology to the 3' end of the DOP primer. The tendency to amplify only those regions flanked by the DOP primer sequence is a significant barrier to the applicability of DOP protocol for amplification of entire genomes.

These findings are in agreement with much of the literature which finds that although DOP-PCR amplified human DNA is sufficient for general genotyping purposes [31], sequence-specific overamplification and subsequent coverage bias results from amplification of smaller size, lower complexity genomes such as cosmids and plasmids [18]. Lower complexity genomes would be expected to have lower, less uniform incidence of the 3' priming sequence than larger genomes, resulting in more bias from the amplified bacterial genomes used our research than the human DNA examined by Cheung *et al.* (1996). Simulated studies of DOP-PCR uniformity found that for a given DOP primer sequence, the specific regions of DOP overamplification could be accurately predicted for several eukaryotic genomes [32], although some (~22%) of the regions expected to amplify in the study failed to generate any product at all, illustrating the additional affect of bias stemming from amplicon amplification efficiency [33]. Dean *et al.* [15] reported a 6-fold bias in coding regions following PCR-based WGA. The increased bias detected by whole genome sequencing in our study might reflect the fact that coding regions have evolved to permit polymerase access and expression, and therefore coding regions may experience less bias than non-coding regions.

Although in our study PEP amplification generated lower bias than DOP, possibly due to increased degeneracy of the PEP primers, both PEP and DOP suffered from potential biases inherent to PCR amplification. PCR-related bias in products amplified from complex mixtures is well documented in the literature, due to factors including differential GC content [34], product reannealing [35] and primer binding energy differences [36]. Moreover, not all amplicons are amenable to PCR amplification [32], resulting in missing sequence, the high-temperature denaturation required to render genomic DNA single-stranded for PCR-based WGA methods can cause cytosine deamination, and homoduplexes can form during the ramp from denaturation to annealing temperature [37]. All of these factors can result in the generation of up to 70% nonspecific amplification artefacts [31] leading to incomplete coverage [38]. This agrees closely with the high percentage of PEP and DOP generated sequences (60 and 80% respectively, as opposed to 40% for unamplified and MDA samples) that failed to map to the reference genomes in our study. This final point is particularly relevant to researchers who are considering the use of WGA amplified material for sequencing. Given the high rate of non-specific or artifactual product formation in the PCR-based WGA methods, the cost of sequencing is increased due to the poor return of sequence data per template. *De novo* sequencing is more drastically affected as the artifactual products might not be removed prior to contig assembly, and thus waste processing time and possibly generate incorrect scaffolds.

Conclusion

In this paper, we utilized whole genome sequencing and sequence-based karyotyping to assess the bias induced by

four methods of whole genome amplification. While the advantages and disadvantages of each WGA method have been discussed extensively in literature [15,16,19,20], previous comparisons between whole genome amplification techniques have used SNP analysis, CGH or FISH analyses to determine induced bias. SNP analysis excels at detecting single base errors, but since SNP analysis typically examines only a small fraction of a given genome, it has the potential to overlook or miss most amplification infidelities. Additionally, as the loci examined by SNP analysis are typically characterized by minimal sequence differentiation, these loci may be relatively immune to amplification bias as their amplification efficiencies would be roughly equivalent. Bias detection via FISH and CGH, on the other hand, is limited to a mapping resolution of approximately 20 Megabases [39,40], is incapable of detecting small scale errors or distortions and has difficulty detecting homozygous deletions [39].

Previously prohibitively expensive and time consuming, whole genome sequencing of entire genomes is now possible with the 454 sequencing system in a matter of days. Using this approach we conducted a high resolution (10 base) examination of amplification-induced bias, detecting statistically significant ($P < 0.001$) bias in all amplified DNA samples relative to unamplified controls. MDA-based amplification methods generated the highest amplification yield and most complete genome coverage while introducing the least bias of all the amplification methods examined. Of the MDA methods, Repli-G generated 3 to 4-fold more amplified DNA, but introduced marginally more bias than GenomiPhi, and generated significantly lower genome coverage when amplifying *Halobacterium* species NRC-1. It is important to note that both MDA based processes were challenged by the high GC content, multi-copy minichromosome containing samples. This may result from the fact that more of the small circular chromosomes are probably present in the starting sample, are more likely than large chromosomes to survive extraction and purification processes intact, and strand displacement is more difficult in regions with increased GC content.

In summary, we have determined that none of the whole genome amplification methods we investigated are free from bias. Depending upon the intended application, researchers should careful weigh the decision whether or not to use whole genome amplification. PCR-based WGA methods generated roughly an order of magnitude less amplified DNA than the MDA methods, with concomitantly increased bias. In our analysis, the relatively low efficiency and yield and high bias generated by the PCR-based methodologies renders them unsuitable for whole genome sequencing and representative amplification of precious DNA. This is not to say that the PCR-based methods are without merit, merely that based on our results, their use should be restricted to DNA amplification for genotyping or marker identification purposes [41], not uniform genomic amplification for high accuracy whole genome sequencing. MDA-based techniques generated significantly higher yields of the proper (non-artifactual) template, and induced lower, but still significant levels of bias. Samples destined for high resolution copy number studies, or complex populations composed of genomes with diverse sizes and GC contents, such as environmental samples, may experience detectable and possibly unacceptable amplification bias. In contrast, applications such as strain identification or whole genome sequencing of purified samples may either tolerate the inherent bias, or surmount it through additional sequence oversampling. For applications amenable to amplified DNA, our investigation determined that MDA-based methodologies produce the greatest amplification yield with the lowest associated bias.

Methods

Template DNA

Bacterial DNA was obtained from two different sources. The *Campylobacter jejuni* was the generous gift from Dr. Jorge Galan (Dept. of Microbial Pathogenesis, Yale University, New Haven CT.). The DNA was subsequently stored at -20°C until used. Ten micrograms of *Halobacterium* species NRC-1 DNA were purchased from ATCC (Order number 700922D, Manassas, VA) and arrived as a lyophilized pellet. The DNA was reconstituted to a concentration of 1 ug/ul in molecular biology grade water (Eppendorf AG, Hamburg, Germany), and stored at -20°C until needed. The genomic DNA served both as template for whole genome amplification methods and as unamplified control DNA. Three microgram aliquots of each bacterial genomic DNA, representing approximately 1.8×10^9 genomic equivalents of *C. jejuni* and 1.1×10^9 genomic *Halobacterium* equivalents, were removed for unamplified controls prior to initiation of the various whole genome amplifications. Once aliquotted, the control DNA samples were stored at -20°C until processed as outlined in Sample Preparation.

Whole genome amplification methods

GenomiPhi

One of the two commercially available Multiple Displacement Amplification kits used in this study was the GenomiPhi DNA Amplification Kit (Amersham Pharmacia, Uppsala, Sweden). For the GenomiPhi reactions, 25 ng of genomic DNA (in 2.5 μl) were mixed with 22.5 μl of GenomiPhi sample buffer. This mix was heat denatured at 95°C for three minutes, then cooled on ice. Twenty-seven microliters of GenomiPhi Reaction buffer were then mixed with 3 μl GenomiPhi enzyme, and 25 μl of this mix were added to the denatured genomic DNA. The reaction

was subsequently incubated at 30°C for 14 hours, then heat inactivated at 65°C for 10 minutes. The amplified DNA was pelleted and purified by sodium acetate precipitation and resuspended in 25 μl of 10 mM Tris (pH 7.5).

Repli-G

The other Multiple Displacement Amplification kit used in this study was the Repli-G Whole Genome Amplification kit (Qiagen Sciences Inc. Germantown, MD). For the Repli-G reactions, 25 ng of genomic material were diluted in TE to a final volume of 2.5 μl. To this was added 2.5 μl of a 1:8 dilution of freshly made Solution A (0.4 M KOH, 10 mM EDTA); the solution was incubated at room temperature after mixing. After three minutes, 5 μl of Stop Solution (1:10 dilution of Repli-G Solution B) were added to the incubated reaction. A solution consisting of 32.4 μl water, 15 μl of Repli-G 4X Mix and 0.6 μl of Repli-G polymerase were mixed, and 40 μl of this solution were added to the 10 μl of denatured, neutralized genomic DNA. The combined solutions were then mixed and incubated at 30°C for 14 hours. The samples were then heat inactivated at 65°C for 10 minutes, and the DNA was pelleted and purified by sodium acetate precipitation and resuspended in 25 μl of 10 mM Tris (pH 7.5).

PEP

For the PEP libraries, *C. jejuni* and *Halobacterium* DNA samples were amplified following the improved PEP protocol (iPEP) [4,5] using the Expand High Fidelity PCR System from Roche Applied Science (Basel, Switzerland). Each 50 μl reaction had a final composition of 0.1 mM dNTPs, 50 μg/ml BSA, 2.5 mM $MgCl_2$, 16 μM of random 15-mer PEP primers (5' NNN NNN NNN NNN NNN 3') from Integrated DNA Technologies (Coralville, IA), 25 ng genomic DNA, 3.5 units of Expand High Fidelity Polymerase, and 1X Expand High Fidelity Buffer. The samples were thermocycled as follows: 2 minute denaturation at 94°C followed by 50 cycles of a 40 second denaturation at 94°C, 2 minutes at 37°C, 0.1 °C/second ramp to 55°C, 4 minutes at 55°C, and 30 seconds at 68°C. A 7 minute extension at 68°C followed the final cycle. After thermocycling, the DNA was pooled and purified using a PCR clean-up kit from Qiagen (Valencia, CA). The Qiagen PCR clean-up kit recovers fragments from 40 bp to 10 kb, thus permitting removal of unextended primers without the loss of amplified fragments.

DOP

For the DOP libraries, *C. jejuni* and *Halobacterium* DNA samples were amplified using the DOP PCR Master Kit from Roche Applied Science. For the reaction, 25 ng of genomic DNA were added to 25 μl DOP PCR Master Mix, 2.5 μl of the supplied 40 μM DOP PCR primer (5' CCG ACT CGA G NNN NNN ATG TGG 3'), with enough water to bring the volume to 50 μl. The manufacturer's instructions were followed for the long thermocycling protocol, except that the reactions were cycled in 50 μl volumes as opposed to the suggested 100 μl volume. The samples were thermocycled as follows 5 minute denaturation at 95°C, followed by 5 cycles of 1 minute at 94°C, 1.5 minutes at 30°C, ramp to 72°C at 0.2°C/second, and 3 minutes at 72°C, then 35 cycles of: 1 minute at 94°C, 1 minute at 62°C, and 2 minutes at 72°C with 14 seconds added to the 72°C step with each cycle, finishing with 7 minutes at 72°C.

Sample preparation and amplification

All samples of genomic DNA, whether amplified or control, were treated identically after purification, following the protocol outlined in Margulies *et al.* [25]. The process is summarized briefly as follows: 3 μg of DNA from the respective sample were fragmented by a five minute nebulization at 44 psi. The DNA was then purified over MinElute columns following the PCR clean-up protocol from Qiagen. The purified, fragmented DNA was polished with T4 DNA Polymerase and T4 Polynucleotide Kinase (New England Biolabs, Beverly, MA). After a half hour reaction with kinase, the DNA fragments were purified again over MinElute columns, before blunt-end ligation (New England Biolabs Quick Ligation Kit) to 454's proprietary double-stranded DNA adaptors A and B. The double-stranded adaptors have one blunt end and one overhanging end, and the blunt end has a non-phosphorylated 5' end, so that the adaptors do not self-ligate. The overhanging 5' end of the B adaptor is biotinylated. The ligated DNA fragments were then bound to magnetic, Streptavidin-coated beads (Dynal Biotech, Oslo, Norway). Because of the non-phosphorylated 5'adaptor ends, the ligated fragments contain nicks, which were displaced with *Bst* polymerase (New England Biolabs). Single-stranded DNA was melted away from the beads with 0.125 M NaOH and the single-stranded fragments were purified over Qiagen MinElute columns. The resulting purified DNA library was run on an Agilent 2100 BioAnalyzer using a RNA 6000 Pico LabChip® (Palo Alto, CA) to quantitate the single stranded DNA concentration, ranging from 200 to 500 bases in length, with a mean size of 350 bases.

The library was combined with a solution containing emulsion PCR reagents and DNA capture beads (Amersham Biosciences NHS-activated HP sepharose) and amplification was carried out as described elsewhere [25]. After amplification, emulsion breaking, and enrichment of the library beads, the non-covalently bound strand was melted away, a sequencing primer was annealed to the covalently bound, amplified strands, and the primed DNA beads were sequenced on the 454 Sequencing System.

Mapping

An alignment algorithm developed by 454 Life Sciences was used to map the reads generated by the 454 sequencing system to the GenBank entries either for *Campylobacter jejuni* (gi|15791399|ref|NC_002163.1| Campylobacter jejuni subsp. jejuni NCTC 11168, complete genome, total length 1.64 Mb, 31% GC) or a composite reference sequence composed of *Halobacterium* species NRC-1 (gi|12057215|ref| AE004437|Halobacterium sp. NRC-1 complete genome, chromosome length 2.01 Mb), and the two associated multicopy minichromosomes pNCR100 (gi|10803547|ref| NC_001869| Halobacterium sp. NRC-1 plasmid pNRC100, complete sequence, 191 Kb in length, 57.9% GC) and pNCR200 gi|12057216ref|AE004438|Halobacterium sp. NRC-1 plasmid pNRC200 complete genome, 365 Kb, 59.2% GC content). This alignment technique matches the light signals measured during each nucleotide flow to the signals expected from the reference sequence. The advantage of this technique is that it takes into consideration that the 454 system sequences DNA one mononucleotide repeat stretch as a time – as opposed to traditional approaches that sequence on a nucleotide-by-nucleotide basis.

Acceptable (or "mapped") alignments were distinguished from rejected (or "unmapped") alignments by calculating the alignment score for each sequence. For this study, an alignment was recorded when the signals from the 5'-end of a read agreed with the expected reference sequence with an average logarithmic probability of at least -1.0. The alignment must also have spanned at least the first 50 light signals from the read. Benchmarking has shown this definition is roughly equivalent to a nucleotide-based alignment with 95% identity over at least the first 30 bases of the read (Data not shown).

Reads that aligned to more than one location on the reference genome were ignored to remove ambiguities concerning the location of the genome from which the DNA actually originated.

Authors' contributions

Both RP and AdW collaborated on the manuscript, and conducted the various methods of whole genome amplification. GJS was responsible for significant amounts of manuscript proofreading and rewriting. MBG was the main source for statistical insight, and ran extensive simulations to suggest which tests should be run for the analysis. KRT conducted all of the DNA sample preparation work, while RNP was responsible for the sequencing. ME and JMR provided invaluable comments and insights into the project, and suggested key directions in which the work should progress. JHL conceived of, and directed the study. All of the nine have reviewed and approved the final manuscript, and have no competing interests to declare.

Additional material

Additional File 1

Table S1. Read distributions between unamplified* Halobacterium *samples. *Kolmogorov-Smirnov comparison of the distributions of reads per bin from an unamplified sample of* Halobacterium *species NRC-1 with four replicate unamplified control libraries. Bin Size refers to the number of bases comprising each individual bin into which the genome was broken for analysis; 100,000 reads were used for each analysis. As no significant differences were found between the distributions, ranked bias values (derived from D statistics) were assumed equivalent and not assigned.* ***Table S2. Read distributions between unamplified* Campylobacter *samples.*** *Kolmogorov-Smirnov comparison of the distributions of reads per bin from an unamplified sample of* Campylobacter jejuni *with four replicate unamplified control libraries. Bin Size refers to the number of bases comprising each individual bin into which the genome was broken for analysis; 100,000 reads were used for each analysis. As no significant differences were found between the distributions, ranked bias values (derived from D statistics) were assumed equivalent and not assigned.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-7-216-S1.doc]

Acknowledgements

The authors would like to acknowledge Dr. Jorge Galan (Dept. of Microbial Pathogenesis, Yale University, New Haven CT.) for his kind gift of the *Campylobacter jejuni* DNA, and Michael S. Braverman (Department of Bioinformatics, 454 Life Sciences) for informative and insightful conversations concerning strategies for comparative genomic evaluation.

454 Life Sciences thanks NHGRI for continued support under grants R01 HG003562 and P01 HG003022 for the development of this platform, as well as all of the enthusiastic employees of 454 who developed the sequencing system

References

1. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E, Williams P, de Chaffoy D, Huitric E, Hoffner S, Cambau E, Truffot-Pernot C, Lounis N, Jarlier V: **A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis.** *Science* 2005, 307:223-227.
2. Ranade K, Chang MS, Ting CT, Pei D, Hsiao CF, Olivier M, Pesich R, Hebert J, Chen YD, Dzau VJ, Curb D, Olshen R, Risch N, Cox DR, Botstein D: **High-throughput genotyping with single nucleotide polymorphisms.** *Genome Res* 2001, 11:1262-1268.
3. Syvanen AC: **Toward genome-wide SNP genotyping.** *Nat Genet* 2005, **37 Suppl**:S5-10.
4. Zheng S, Ma X, Buffler PA, Smith MT, Wiencke JK: **Whole genome amplification increases the efficiency and validity of buccal cell genotyping in pediatric populations.** *Cancer Epidemiol Biomarkers Prev* 2001, 10:697-700.
5. Dietmaier W, Hartmann A, Wallinger S, Heinmöller E, Kerner T, Endl E, Jauch KW, Hofstädter F, Rüschoff J: **Multiple mutation analyses in single tumor cells with improved whole genome amplification.** *American Journal of Pathology* 1999, 154:83-95.
6. Hahn S, Zhong XY, Troeger C, Burgemeister R, Gloning K, Holzgreve W: **Current applications of single-cell PCR.** *Cell Mol Life Sci* 2000, **57**:96-105.

7. Lasken RS, Egholm M: **Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens.** *Trends Biotechnol* 2003, 21:531-535.
8. Schloss PD, Handelsman J: **Metagenomics for studying unculturable microorganisms: cutting the Gordian knot.** *Genome Biol* 2005, 6:229.
9. Raghunathan A, Ferguson HRJ, Bornarth CJ, Song W, Driscoll M, Lasken RS: **Genomic DNA amplification from a single bacterium.** *Appl Environ Microbiol* 2005, 71:3342-3347.
10. Cowan DA: **Microbial genomes--the untapped resource.** *Trends Biotechnol* 2000, 18:14-16.
11. Rothwer F, Seguritan V, Choi DH, Segall AM, Azam F: **Production of shotgun libraries using random amplification.** *Biotechniques* 2001, 31:108-12, 114-6, 118.
12. Wang G, Maher E, Brennan C, Chin L, Leo C, Kaur M, Zhu P, Rook M, Wolfe JL, Makrigiorgos GM: **DNA amplification method tolerant to sample degradation.** *Genome Res* 2004, 14:2357-2366.
13. Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S: **Ancient DNA.** *Nat Rev Genet* 2001, 2:353-359.
14. Paabo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M: **Genetic analyses from ancient DNA.** *Annu Rev Genet* 2004, 38:645-679.
15. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS: **Comprehensive human genome amplification using multiple displacement amplification.** *Proc Natl Acad Sci U S A* 2002, 99:5261-5266.
16. Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, Dillon D, Segraves R, Vossbrinck B, Gonzalez A, Pinkel D, Albertson DG, Costa J, Lizardi PM: **Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH.** *Genome Res* 2003, 13:294-307.
17. Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N: **Whole genome amplification from a single cell: implications for genetic analysis.** *Proc Natl Acad Sci U S A* 1992, 89:5847-5851.
18. Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BJ, Tunnacliffe A: **Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer.** *Genomics* 1992, 13:718-725.
19. Wells D, Sherlock JK, Handyside AH, Delhanty JD: **Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation.** *Nucleic Acids Res* 1999, 27:1214-1218.
20. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M, Lasken RS: **Unbiased whole-genome amplification directly from clinical samples.** *Genome Res* 2003, 13:954-964.
21. Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvanen AC: **Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA.** *Nucleic Acids Res* 2003, 31:e129.
22. Bannai M, Higuchi K, Akeda T, Furukawa M, Yamaoka M, Sato K, Tokunaga K: **Single-nucleotide-polymorphism genotyping for whole-genome-amplified samples using automated fluorescence correlation spectroscopy.** *Anal Biochem* 2004, 327:215-221.
23. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, Li C, Meyerson M, Sellers WR: **Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification.** *Nucleic Acids Res* 2004, 32:e71.
24. Ng WL, Kothakota S, DasSarma S: **Structure of the gas vesicle plasmid in Halobacterium halobium: inversion isomers, inverted repeats, and insertion sequences.** *J Bacteriol* 1991, 173:1958-1964.
25. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, 437:376-380.
26. Yap EP, McGee JO: **Short PCR product yields improved by lower denaturation temperatures.** *Nucleic Acids Res* 1991, 19:1713.
27. Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC: **Mutation detection and single-molecule counting using isothermal rolling-circle amplification.** *Nat Genet* 1998, 19:225-232.
28. Agarwal RK, Perl A: **PCR amplification of highly GC-rich DNA template after denaturation by NaOH.** *Nucleic Acids Res* 1993, 21:5283-5284.
29. Dean F, Nelson J, Giesler T, Lasken R: **Rapid amplification of plasmid and phage DNA using Phi29 polymerase and a multiply-primed rolling circle amplification.** *Genome Research* 2001, 11:1095-1099.
30. Abulencia CB, Wyborski DL, Garcia JA, Podar M, Chen W, Chang SH, Chang HW, Watson D, Brodie EL, Hazen TC, Keller M: **Environmental whole-genome amplification to access microbial populations in contaminated sediments.** *Appl Environ Microbiol* 2006, 72:3291-3301.
31. Cheung V, Nelson S: **Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA.** *Proc Natl Acad Sci USA* 1996, 93:14676-14679.
32. Jordan B, Charest A, Dowd JF, Blumenstiel JP, Yeh RF Rf, Osman A, Housman DE, Landers JE: **Genome complexity reduction for SNP genotyping analysis.** *Proc Natl Acad Sci U S A* 2002, 99:2942-2947.
33. Kwok PY: **Making 'random amplification' predictable in whole genome analysis.** *Trends Biotechnol* 2002, 20:411-412.
34. Benita Y, Oosting RS, Lok MC, Wise MJ, Humphery-Smith I: **Region-alized GC content of template DNA as a predictor of PCR success.** *Nucleic Acids Res* 2003, 31:e99.
35. Suzuki MT, Giovannoni SJ: **Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR.** *Appl Environ Microbiol* 1996, 62:625-630.
36. Polz MF, Cavanaugh CM: **Bias in template-to-product ratios in multitemplate PCR.** *Appl Environ Microbiol* 1998, 64:3724-3730.
37. Kurata S, Kanagawa T, Magariyama Y, Takatsu K, Yamada K, Yokomaku T, Kamagata Y: **Reevaluation and reduction of a PCR bias caused by reannealing of templates.** *Appl Environ Microbiol* 2004, 70:7545-7549.
38. Paunio T, Reima I, Syvanen AC: **Preimplantation diagnosis by whole-genome amplification, PCR amplification, and solid-phase minisequencing of blastomere DNA.** *Clin Chem* 1996, 42:1382-1390.
39. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nature Genetics* 1999, 23:41-46.
40. Gray JW, Collins C: **Genome changes and gene expression in human solid tumors.** *Carcinogenesis* 2000, 21:443-452.
41. Hawkins TL, Detter JC, Richardson PM: **Whole genome amplification--applications and advances.** *Curr Opin Biotechnol* 2002, 13:65-67.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral